SENTIMENT ANALYSIS OF PATIENT FEEDBACK

by

PHILLIP SMITH

A thesis submitted to the

University of Birmingham

for the degree of

DOCTOR OF PHILOSOPHY

School of Computer Science

College of Engineering and Physical Sciences

University of Birmingham

December 2015

Abstract

The application of sentiment analysis as a method for the automatic categorisation of opinions in text has grown increasingly popular across a number of domains over the past few years. In particular, health services have started to consider sentiment analysis as a solution for the task of processing the ever-growing amount of feedback that is received in regards to patient care. However, the domain is relatively under-studied in regards to the application of the technology, and the effectiveness and performance of methods have not been substantially demonstrated.

Beginning with a survey of sentiment analysis and an examination of the work undertaken so far in the clinical domain, this thesis examines the application of supervised machine learning models to the classification of sentiment in patient feedback. As a starting point, this requires a suitably annotated patient feedback dataset, for both analysis and experimentation. Following the construction and detailed analysis of such a resource, a series of machine learning experiments study the impact of different models, features and review types to the problem. These experiments examine the applicability of the selected methods and demonstrate that model and feature choice may not be a significant issue in sentiment classification, whereas the type of review that the models train and test across does affect the outcome of classification. Finally, by examining the role that responses play in the patient feedback process and developing the idea of incorporating the inter-document context provided by the response into the feedback classification process, a recalibration framework for the labelling of sentiment in ambiguous texts for patient feedback is developed.

As this detailed analysis will demonstrate, while some problems in performance remain despite the development and implementation of the recalibration framework, sentiment analysis of patient feedback is indeed viable, and achieves a classification accuracy of 91.4% and F_1 of 0.902 on the gathered data. Furthermore, the models and data can serve as a baseline to study the nature of patient feedback, and provide a unique opportunity for the development of sentiment analysis in the clinical domain.

I dedicate this thesis to

the memory of my father,

Robert H. Smith

(1946-2010)

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor, Mark Lee, for the support, guidance and wisdom that he has imparted during my studies at the University of Birmingham. From academia to industry, and back again, Mark has had an extraordinary influence on my development as a doctoral researcher, and for this, I am truly grateful. I must also thank John Barnden and Peter Hancox for the encouraging feedback that they gave during our thesis group meetings.

In relation to the content of this thesis, I am grateful to the UK government for providing access to the NHS Choices patient feedback data under the Open Government Licence[1]. I am also grateful to the School of Computer Science for funding my doctoral research through a Ph.D. Scholarship.

I would like to acknowledge my friends and colleagues at the University of Birmingham for creating such a wonderful environment within which to study. In particular, I would like to thank Sheng Li for the lively discussions we have had over the years regarding the nature of sentiment, in its many forms.

It's difficult to say where I would be today if it weren't for the incredible encouragement that my family has given, and for the sacrifices that have been made. You have given me the strength to carry on through some tough times, so thank you.

Finally, I would like to thank Sarah for helping me turn that corner and see things in a way that no other could. I am, and always will be, deeply thankful for the patience, care and love she has not only given to me, but also to our son, Rupert.

[1] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF PUBLICATIONS

- Phillip Smith & Mark Lee (2012) "Cross-discourse Development of Supervised Sentiment Analysis in the Clinical Domain", in *Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*. Association for Computational Linguistics, pages 79-83, Jeju Island, Korea.

- Phillip Smith & Mark Lee (2013) "A CCG-Based Approach to Fine-Grained Sentiment Analysis in Microtext", in *AAAI Spring Symposium Series: Analyzing Microtext (SAM)*, pages 80-86, Palo Alto, California, USA.

- Phillip Smith & Mark Lee (2014) "Acknowledging Discourse Function for Sentiment Analysis", In Alexander Gelbukh (ed.): *Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*. Springer , Lecture Notes in Computer Science, 8404(2), Springer, pages 45-53, Kathmandu, Nepal.

- Phillip Smith & Mark Lee (2015) "Sentiment Classification via a Response Recalibration Framework", in *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*. Association for Computational Linguistics, pages 175-180, Lisboa, Portugal.

CHAPTER 1

INTRODUCTION

1.1 Sentiment analysis in healthcare

Healthcare is an important, but sensitive, aspect of our everyday well-being. Over the course of our lives, we will undoubtedly have some form of interaction with a health service. While common ailments can be treated in a prescribed manner with uniform processes, healthcare is a highly personal service. Dependent on condition and the resources available to treat it, the outcome of a patient's experience with a health service can greatly vary. A treatment can go extremely well and a high standard of care can be received or a treatment can go very badly, and the standard of care may be below what is reasonably expected. Ideally, a high standard of care would be preferable, but there are many factors, some unknown, that can affect how a patient reacts to their experience with a health service provider. Thankfully, given good or bad patient experiences, there are a variety of mechanisms in place to handle and collect their feedback about the health service.

The role of patient feedback in a publicly funded health service such as the NHS is vital. The comments that are received are not only important in boosting the morale of all staff involved in the care-giving process, but also for indicating where improvements need to be made to enhance the standard of care that is provided.

In 2012, the Prime Minister of the United Kingdom announced a method to improve patient care in England (GOV.UK, 2012): the Friends and Family test (FFT). This was a simple metric,

based upon a single question that was asked to determine whether a patient would recommend the health service to their friends and family if they required similar care. This test was implemented in a number of services provided by the NHS, including, maternity wards, dentists and accident and emergency departments across the UK. Having initiated this mechanism in April 2013, by February 2015 the FFT had received five million responses (NHS England, 2015). While this only covers a fraction of all patients that were treated in the NHS over that period, this is still a substantial amount of feedback regarding a single service.

The responses that were collected indicated on a five-point scale how likely a patient was to recommend the service, ranging from extremely unlikely to extremely likely. Given the resulting recommendation, some trusts asked a follow-up question to find out further details about what in particular a patient liked and disliked, and if they had any advice to give. However, this follow-up question was not a requirement to the test, and not all trusts implementing the FFT metric followed up on the initial response in order to request specific details about a patient's opinion.

Where follow-up questions were asked and collected, some trusts used sentiment analysis software to identify trends in the resulting verbatim feedback (NHS England, 2014). This was not used by all trusts however, as some collected responses using paper-based feedback, which required manual transcription before analysing the feedback. As transcription of handwritten text is a relatively time-consuming job, often requiring many transcribers to interpret and manually input the verbatim comments into a database, this limited the ability for the software to analyse some of the comments.

Despite a number of the involved trusts using software to quantitatively examine the verbatim comments, the review document of the FFT (NHS England, 2014) indicated that at a high level sentiment analysis was being treated with heed, as the effectiveness and accuracy of the methods in the clinical domain had not been proven. This is understandable given the fairly limited work carried out examining the classification of sentiment in documents from the clinical domain. However, as society communicates in an ever increasing digital manner, and with the expectation that the NHS will experience a growth in the number of online comments it

receives, the task of manually analysing each verbatim item of feedback becomes increasingly more challenging. Automated sentiment analysis provides one potential solution to this issue, but this field is still developing and maturing, and underlying issues exist that limit the performance of sentiment classification algorithms. One of the major overlapping research questions between the work in sentiment analysis and that in the clinical domain is to what extent can sentiment analysis reliably be carried out on a document set of patient feedback? In this thesis we address this overarching research question by thoroughly examining a number of related sub-questions to determine the level of applicability of sentiment analysis to the patient feedback domain.

1.2 Context in sentiment analysis

State-of-the-art approaches to the classification of clinical documents into their respective sentiments have focused on the investigation of the application of a variety of machine learning methods for the given task (Greaves et al., 2013; Xia et al., 2009). Results of these studies show promise, yet all admit that if the state of the art is to be improved, then context must be considered in the classification process, particularly in the medical domain (Denecke & Deng, 2015).

Context is important in sentiment analysis as it can be used to clarify the sentiment of a document. For example, where a sentiment is conveyed implicitly, perhaps through the use of objective medical terminology regarding the health status of a patient or the side effects of a treatment, then the context of the document can be used to guide the classification process and suggest that there is a sentiment present within its contents.

Context can often be gleaned by examining the company a word keeps (Firth, 1957), but sometimes this does not fully resolve the problem nor obtain the context that is necessary to adequately determine the sentiment conveyed in a document. Denecke & Deng (2015) suggest that to fully interpret the sentiment articulated by documents in the medical domain, context should be considered beyond the boundary of the document. Currently, examined systems do

not make this consideration as machine learning techniques applied to sentiment classification to yield state-of-the-art results in the clinical domain only regard a review in isolation during the classification process, hence limiting the scope of the classification.

In this thesis, we examine the potential for introducing an external and related context to the classification process that goes beyond the document boundary. The nature of current datasets limits the ability to use related documents to give context to a particular document, and consequently use such contextual information to recalibrate classifier output, as typically reviews don't tend to have relevant, context-bearing documents given in close proximity. We examine the case where a context can be introduced through a healthcare provider's response to an instance of patient feedback, and investigate the extent to which the classifier output can be recalibrated where the initial document classification may be incorrect, with the intention of producing a more accurate and robust standard of sentiment classification in the clinical domain.

1.3 Research questions

Patient opinion is an informal yet valuable metric that can be used to determine the standard and quality of care provided by a health service. However, to analyse any type of feedback requires a time-consuming manual examination to read and aggregate the viewpoints of many different patients to establish the aspects of the patient experience that require improvement.

This thesis will focus on the use of supervised machine learning techniques to automate the process of accurately classifying an item of patient feedback by the sentiment that it conveys.

A dataset of real-world patient feedback, the NHS Choices Sentiment Dataset (NCSD), is compiled to train the supervised machine learning approaches about the way that sentiment is conveyed in patient feedback. This dataset is also used as a test bed to evaluate the applicability of the examined machine learning approaches. In the NCSD, the instances of patient feedback data are split into four separate documents that are all linked by a unique comment identifier: verbatim comments from the patient detailing their likes, dislikes and advice, and a correspond-

ing response to the review from the health service provider. The NCSD is unlike other sentiment analysis datasets that typically contain a collection of reviews, consisting of only a single review field (Pang et al., 2002; Blitzer et al., 2007; Maas et al., 2011). Such reviews may still contain the likes, dislikes and advice of an author intermingled within the course of a review document. However, by having these respective elements of a reviewer's opinions explicitly separated from one another, a classifier can be trained on the explicit positive and negative items of feedback, and any sentiment that is conveyed in the advice.

The separator between the fields therefore poses an interesting theoretical issue in the domain of sentiment analysis: namely, how is sentiment communicated and does the type of document make a difference when training a supervised machine learning classifier. Both form part of an item of feedback, but the split suggests that the sub-documents may be requested for different purposes. The likes and dislikes are focused descriptions of a patient's high-level experience with the health service that clearly are associated with positive or negative sentiments. The advice, on the other hand, is less constrained in the sentiment that should be conveyed in this field, and often the advice is an extension of the likes and dislikes, elaborating on the salient aspects of the experience. Three review types can be distinguished from the separated fields: (a) those containing only likes and dislikes (Type 1), (b) those containing only free-form text giving a summary or advice (Type 2), and (c) a combination of likes, dislikes and advice (Type 3). With these in mind, over the course of this thesis the applicability of these different types of review document for sentiment analysis is examined, and the question of whether one type of review is able to generate a model of sentiment in user reviews that is better than the others. Thus far, state-of-the-art methods developed for the sentiment classification of clinical documents have not examined the effect of different review types. This therefore leads to the construction of the first research question examined in this thesis:

- To what extent does review type affect the outcome of the sentiment classification of patient feedback?

The Type 1 and Type 2 reviews can be seen as serving two distinct purposes related to the discourse function of the texts. Type 1, conveying the *likes* and *dislikes* of the patient can be

viewed as performing an expressive discourse function. Type 2 on the other hand, a less constrained field labelled *advice*, can be seen as attempting to change aspects of the health service, and can be seen as conveying a persuasive discourse function. The two distinctly convey a sentiment however, yet perform different acts in their utterance. This therefore leads to the question of whether the different discourse functions have an effect on the overall classification performance. Again, state-of-the-art methods for the sentiment classification of clinical documents have not examined the affects of training and testing across review type. This therefore leads to the second research question that we investigate in this work:

- To what extent does training a model on a document set that serves a different, but related, purpose to that of the test data affect the performance of supervised machine learning classifiers trained for the sentiment classification of patient feedback?

The type of language used in the review is not the only factor that can affect the outcome of sentiment classification; of course, the choice of supervised machine learning algorithm will have a role in the outcome of classification. Throughout the literature, different supervised machine learning approaches have been applied to the problem of sentiment classification, across a number of different domains (Bobicev et al., 2012; Blitzer et al., 2007; Pang et al., 2002). This has lead to variations in the suggested suitability of classification algorithms. In comparison to social media, or film and product reviews, patient feedback is a field that has not received as much focus within the sentiment classification literature. Therefore, the applicability of classification algorithms that perform well in other domains for sentiment analysis must also be studied in the scope of this study to ascertain their applicability. This leads to the third research question that will be answered by this thesis:

- To what extent does classifier choice affect the outcome of the sentiment classification of patient feedback?

Following on from this question, there is another factor that tends to limit the performance of machine learning classifiers applied to the problem of sentiment classification in text: the

choice of feature upon which the classifier is trained and tested. As we are dealing with discrete linguistic data, the feature choice requires a conversion process into a form that a classification algorithm is able to use. Many potential features have been proposed in the machine learning literature which contribute to classifier performance (Sebastiani, 2002). It is unclear which may be most suited to the choice of classifier and also the classification of sentiment in clinical documents. This generates the fourth research question that this thesis will examine:

- To what extent does the choice of feature representation affect the outcome of the sentiment classification of patient feedback?

The first four questions focus on the type of review data, the classifier choice and the feature choice when examining the sentiment of a document in the clinical domain. However, there are additionally prevalent problems that are of a more implicit nature in sentiment classification: namely how to deal with language that does not convey sentiment in an explicit manner. In examples of this implicit usage, words that are typically representative of a particular sentiment are used in unconventional ways in order to introduce a different meaning to that which appears to be presented by the lexical items of a review, and therefore, the review conveys an unexpected sentiment. This covers instances of metaphor, humour and irony in language that current systems struggle to deal with when the domain is not limited or constrained (Maynard & Greenwood, 2014).

These problems typically rely on huge knowledge bases (Schulder & Hovy, 2014) in order to cope with the chance occurrence of these linguistic structures. However, one potential clue to the use of these constructs is the context within which they are used. Context can be internal or external to a document, and where a context forms part of an implicit dialogue, which is the case for reviews, then a response or reply to the comment is able to act as a source of context for the given review. The NCSD does not only contain structured patient feedback, but also an organisation response from the health service. By responding to what a reviewer has written, through the wording of their reply, the responder reveals insights into the sentiment of the original review. The sentiment that is echoed in the response is of importance to the classification process, and the approach that is developed in this thesis is to our knowledge one

of the first to examine the use of review responses in a sentiment classification framework. To test the hypothesis that a review response can be used to recalibrate the outcome of the sentiment classification of patient feedback, we construct the final research question that will be examined in this thesis:

- To what extent can a review response be used to improve the classification results of a collection of patient feedback?

1.4 Methodology

Much of the work in this thesis is discussed with reference to the development of a functional sentiment classification system. The system provides a generic structure for classifying multiple instance based document sets by the sentiment of a specified document, and while tested on patient feedback data, is not at all restricted to the classification of documents in this domain. The system was developed using the Java programming language and used a number of off the shelf libraries to aid the classification process.

The first set of experiments explores the choice of classification model, feature, and data type for classifying patient feedback data by the sentiment that it conveyed. This component is developed to learn from a training set, and apply the learned model to classify an unseen test set. The system incorporates a cross-validation procedure to ensure consistency in results across all possible data. Evaluation of the variables is carried out with respect to the NCSD.

The second set of experiments examines the role of professional responses in recalibrating the output of an initial supervised sentiment classifier. Again, this uses the NCSD dataset to evaluate our proposed approach as to our knowledge no other comparable datasets have been made available for public use.

1.5 Contributions

Given the research questions, this thesis makes the following contributions to knowledge:

1. When classifying patient feedback by the sentiment that it conveys, choice of supervised machine learning model is not a significant issue. Extensive experimentation on a number of models does not yield one particular classification model that is significantly better than the others, although the Multinomial Naïve Bayes classification model consistently ranks as one of the best performing models.

2. When classifying patient feedback by the sentiment that it conveys, choice of document representation and feature weighting for training and testing the supervised machine learning models is not a significant factor, although the use of lower-case stemmed words is consistently ranked as one of the best textual feature representations used.

3. The type of review used for classifying the sentiment of patient feedback is not a significant issue. Classification models trained and tested on reviews of the same type tend to be able to classify sentiment to a comparable degree across all review types. A comparable degree of classification can also be achieved by only considering the final sentences of a review in the classification process.

4. When training and testing *across* review type the overall performance of sentiment classification is negatively affected. In particular, if we associate the review types with specific functions of discourse, the results of the experiments in this thesis indicate that training and testing across datasets with differing discourse functions is detrimental to classification performance.

5. Given the presence of a contextual document, a novel approach is developed whereby the outcome of the sentiment classification of patient feedback is able to be recalibrated with significant increases in overall classification performance.

The development of these contributions has led to the following resulting products:

1. A recalibration system that increases the performance of sentiment classification in text given a contextual document using a probabilistic thresholding procedure.

2. A dataset of patient feedback for sentiment analysis annotated with both polarity and review type information.

3. An annotated dataset of organisational responses to patient feedback, which is used in the recalibration framework.

1.6 Thesis structure

This chapter has presented the motivation for this thesis and the methodology that has been used. The remainder of this thesis is organised as follows:

Chapter Two discusses the different approaches to sentiment analysis, and how the development of the field has branched off to tackle the task of sentiment classification in the clinical domain. This chapter lays the groundwork for the thesis and the motivation for approaching the task of classifying patient feedback from a supervised machine learning perspective.

With this in mind, Chapter Three describes the development of the NCSD, details the annotation process, and finally, gives a corpus analysis of the patient feedback. The analysis confirms the suitability of the data for sentiment classification, and in Chapter Four experiments to examine the applicability of different models, features and review types to the classification of sentiment in patient feedback. This highlights the ability to classify the sentiment of such text, but the issue of implicit and contextual polarity still remains. In Chapter Five, we discuss the approaches that have been developed to tackle the problem of context in sentiment classification, and we propose a framework to recalibrate classifier outcome given a relative context, discussed in Chapter Six. Finally, in Chapter Seven, we conclude and describe avenues for future work.

CHAPTER 2

LITERATURE REVIEW

Introduction

In this thesis, we consider the application of computational sentiment analysis techniques to patient feedback. In forming this literature review, relevant books, theses, journals, and the proceedings of conferences and workshops in the computational linguistics and machine learning domains were examined, with a particular focus on the work on the topic of sentiment classification. Resulting from this, in this chapter, a review of the state of the art in sentiment analysis is given, detailing the different methodological approaches developed to tackle the problem. The review of these techniques concludes with a discussion of the drawbacks experienced by current methodologies, which in turn, motivates the remainder of the thesis. The final section of the chapter then gives a review of the work on sentiment analysis that has been specifically applied to the clinical domain. This section discusses the literature from two perspectives, the classification of biomedical texts, and the classification of patient feedback. It is the latter sub-field of the domain that is the focus of this thesis.

2.1 Sentiment Analysis

This thesis is concerned with the task of sentiment analysis, sometimes referred to by the moniker, opinion mining (Liu, 2015). Sentiment analysis is a task derived from text classifi-

cation, so by way of inheritance, sentiment analysis follows a similar problem scope: given a set of documents, assign each document in the set to an appropriate category that reflects the sentiment conveyed. While text classification may answer the question 'what topic is mentioned in this document?', sentiment analysis answers the question 'what attitude is conveyed in this document?' Attitude is a type of affective state, defined by Scherer (1984) as the "*relatively enduring, affectively coloured beliefs, preferences, and predispositions towards objects or persons*". Examples of attitudes given by Scherer are *liking, loving, hating, valuing* and *desiring*. These can be grouped into higher-level, relative sentiments: positive and negative.

This ability to classify a document by the sentiment it conveys is desirable in a number of different application settings. In particular, it is desirable for the activity of decision making. For example, if we are faced with the choice of buying a new camera, yet are not an expert, then specialist advice can help us come to a decision. As businesses have left the high-street, the ability to talk to someone about buying a new camera has somewhat shrunk. However, as e-commerce websites, such as Amazon, have matured, they have adapted to the needs of the consumer by providing the ability for customers to leave reviews regarding a product they have purchased. Faced with the choice of many cameras, and many more reviews per camera, digesting and understanding the reviews of all the cameras in order to make a decision about which camera is most suitable for you can be a mammoth task. Seminal work in sentiment classification looked to solve this problem by providing a mechanism for extracting the pros and cons from online reviews and giving a succinct overview (Liu et al., 2005).

Looking at this example from a different angle, companies that are either currently selling cameras or seeking to produce a new camera can use a similar sentiment analysis process to find out what a subset of consumers like and dislike about aspects of cameras that are currently on the market. Previously, a time consuming process was required to find this information out through in-depth consumer surveys, however now this is a matter of computational power and presentation techniques. This process can go beyond the analysis of the views of a group of people about certain products, to examine the general sentiment of movie reviews (Scheible & Schütze, 2013), the stock market (Nguyen & Shirai, 2015) and even attempt to predict the

outcome of elections (Chung & Mustafaraj, 2011). Where there is an extractable opinion from a textual data source, there is the ability to apply sentiment analysis techniques to evaluate the sentiment that it conveys.

To enable these applications to come to fruition, the internet has been an important driving force. Previous research in computational linguistics focussed on the development of specifically constrained linguistic problems whereby an algorithm would work for a developed set of examples. Expanding such examples to larger test sets would be difficult due to the difficulty of data distribution. The internet has alleviated this problem. Web crawlers have helped to develop datasets, and increasing network speeds have aided in corpus distribution. It is the relative amount of content that is generated daily that has contributed to the uptake of larger and more complex corpora. This constant flow of data through social networks, such as Twitter, has pushed forward the development of datasets and enabled the progression of approaches to sentiment analysis. Approaches to sentiment classification can at a high level broadly be grouped into approaches that rely on the development and study of the sentiment lexicon such as early work in the field developed by Hatzivassiloglou & McKeown (1997) and those that attempt to learn patterns of sentiment expression from annotated data, which can be traced back to early work such as that of Pang et al. (2002) and Dave et al. (2003).

Many approaches seek to classify the sentiment at the document level, that is, what sentiment the document in its entirety is conveying. However, at a lower, more fine-grained level, are the works in sentiment analysis that aim to determine the opinion holder in a text (Bethard et al., 2006), or the identification of the targets of opinions in a text, often referred to as aspect-based sentiment analysis (Pontiki et al., 2015). Although both the high and low-level tasks differ in their granularity of classification, at their core is the data used for classification and the categories of sentiment that classification algorithms use to guide the sentiment analysis process.

2.1.1 Categories of sentiment

The most common approach to sentiment classification is to categorize a document into a single category from the set $\{positive, negative\}$. This is often referred to as the binary sentiment categorisation problem (Pang et al., 2002). However, this implies that there must be a polarity associated with a document, which is not always the case. Due to this, a third category, *neutral*, is sometimes included to make sentiment classification a multi-class classification problem (Hu et al., 2013b; Koppel & Schler, 2006). However, the *neutral* category has been found to be a highly problematic category to classify a document into, as differentiating between a document conveying a neutral opinion and one that does not convey an opinion at all can be difficult to distinguish (Kim & Hovy, 2004). Some works in the literature also include the category of sarcasm amongst the basic categories as an orientation of sentiment (Maynard & Greenwood, 2014; Liebrecht et al., 2013; Pustejovsky & Stubbs, 2012), although this is a questionable and potentially skewed category for classification due to the implied contempt that may be present.

Beyond a coarse-grained labelling scheme such as the binary described above, efforts have been made to annotate and classify the sentiment of a document set on a fine-grained scale in an attempt to aid the ranking or comparison of multiple reviews (Pang & Lee, 2005; Snyder & Barzilay, 2007). Using a scale from +5 for extremely positive to -5 for extremely negative, with a labelling of 0 being neutral, Taboada et al. (2011) annotate the adjectives, adverbs, nouns and verbs of a multi-domain review corpus to construct a sentiment lexicon. While initially admitting this fine-grained scale is somewhat arbitrary, inter-annotator agreement and dictionary comparison studies validate the robustness of such a scale. One particular benefit of such a fine-grained scale for sentiment annotation is where the polarity of a word is altered by a negator. In the coarse-grained case, this would simply flip the polarity, but this lacks the subtle effects of a negating word. Taboada et al. (ibid.) describe the effects of negating the word *excellent*. This is initially labelled as +5, but preceding this with *not* when using the flipping approach would alter the polarity to -5, which does not seem consistent with other words labelled as extremely negative, such as *atrocious*. The notion of shift negation is able to be applied in such a fine-grained labelling scheme, whereby a valence is shifted by 4 instead, thereby preserving

the nuances of the negator.

Related categorisation schema

In addition to the sentiment classes that have been applied to the classification problem, there are several related categorisation schema that have also been proposed that have been shown to be both relevant and potentially beneficial to sentiment analysis. The first is the classification of documents as subjective or objective (Volkova et al., 2013; Abdul-Mageed et al., 2011; Riloff & Wiebe, 2003). Some see this as a precursor to the sentiment classification problem, as linguistic expressions of sentiment are typically personal to the writer, and exhibit the speaker's private state, so by filtering out objective parts of a document, the task of sentiment analysis is supposedly simplified (Wiebe et al., 2004). However, Scheible & Schütze (2013) have argued that this notion is not optimal for sentiment analysis, as although objective statements do not directly exhibit the private state of an author, they may still have an associated sentiment due to the connotational qualities of the content.

Datasets for sentiment classification

Given the choice of class structure for sentiment classification, the next important consideration is the data that will either be used as a starting point to develop a sentiment lexicon or to train a supervised machine learning model. Datasets differ from work to work, however there are some which tend to be popular in the literature sentiment analysis, and are almost defining of the nature of the task in the scope of their respective domains. These datasets can be grouped into those containing consumer reviews of products (McAuley & Leskovec, 2013b; Jindal & Liu, 2008; Blitzer et al., 2007), films (Pang & Lee, 2005) and restaurants (Ganu et al., 2009); stock market analysis data (Bollen et al., 2011), news headlines (Strapparava & Mihalcea, 2007), meeting dialogues (Somasundaran, 2010), political debates (Carvalho et al., 2011; Thomas et al., 2006) and Twitter data (Ghosh et al., 2015; Li, 2014; Maynard & Greenwood, 2014), amongst others.

2.2 Computational approaches to sentiment classification

Sentiment analysis can be formalised as a classification task, and in doing so, machine learning methods can be applied as a potential computational solution to the problem. A subset of machine learning methods make use of training data to learn how sentiment is expressed, which then leads to the formation of classification parameters and rules in a model that can be applied to unseen documents in order to classify them by the sentiment that they convey. This approach is referred to as the supervised approach.

The opposing approach that can be used to classify the sentiment of a document does not learn directly from training data, but instead, employs the use of an already formed knowledge base to identify the presence of sentiment-bearing words in a text and consequently classifies a document using a method that takes into account the information provided by the resource. This type of approach can often be referred to as the unsupervised approach (Liu, 2012), and the resource that forms the basis for classification using this method is a sentiment lexicon.

2.2.1 Unsupervised approaches

Hu et al. (2013a) suggest that unsupervised approaches to sentiment classification are preferable due to the expense of acquiring sentiment labels for unlabelled data. Due to this, they define the unsupervised method as one that relies on a pre-defined sentiment lexicon to calculate the sentiment conveyed in a given document. Therefore, a focus of the unsupervised approach to sentiment classification is the construction of a suitable lexical resource for the task.

Sentiment lexicon compilation

A lexicon in this work is defined as a digital file containing a vocabulary of the language. It follows then that a sentiment lexicon is a digital file containing the vocabulary of words and phrases that each have an associated polarity. Work in sentiment analysis has led to the generation of a number of sentiment lexicons spanning a number of domains, and the process of sentiment classification using these resources is dependent upon the presence or absence of

words or phrases from the lexicon's vocabulary in a document. The general approach to using the lexicon in classifying a document is to use it as a point of reference, and given this resource, calculate the sum of the sentiment bearing words or phrases in a document in order to label the document with the resulting predominant sentiment, as described in the work of Palanisamy et al. (2013), for example. The application of the resource will be discussed further in the following subsection.

The main differences that can be found between sentiment lexicons are the number of terms that they contain, the variety of terms that they contain, and the level of detail that they give in regards to the sentiment of a particular word in a given lexicon. For example, some lexicons focus on compiling a list of words with certain parts-of-speech, such as adjectives (Hatzivassiloglou & McKeown, 1997), some focus on only those words that appear to be subjective and polarity inducing, irrespective of part-of-speech (Wilson, 2008b), and others attempt to cover the whole vocabulary of a language (Esuli & Sebastiani, 2006). It is the varying methods of generation that leads to these differences between sentiment lexicons.

Three approaches to building sentiment lexicons have been discussed in the literature: manual lexicon derivation using human annotators (Stone, 1966; Wilson & Wiebe, 2005; Pennebaker et al., 2007), dictionary-based methods that use a lexical database such as WordNet (Miller, 1995), whereby words are grouped into sets of cognitive synonyms that can be traversed to form a sentiment lexicon (Baccianella et al., 2010), and finally, corpus-based methods, that given a set of seed terms, will explore the surrounding context of the seed's usage within a corpus to determine additional terms that may share its sentiment (Lu et al., 2011; Zhang & Liu, 2011).

The corpus-based approach is a data driven procedure for sentiment lexicon generation that observes and considers the word distribution in a number of relevant corpora when building a sentiment lexicon. One of the earliest approaches that used a corpus-driven approach to sentiment lexicon development was undertaken by Hatzivassiloglou & McKeown (1997), who constructed a system that observed conjunctions of adjectives in a corpus to determine their respective semantic orientation. They tested their approach to lexicon development using the

Wall Street Journal corpus to extract a set of adjectives that appeared 20 times, removing those with no semantic orientation, and labelling the remainder as positive or negative. This created a seed set of 657 and 679 positive and negative adjectives respectively which was used to search for conjoined sentiment bearing adjectives that while not as frequent, were still potentially sentiment-bearing. The concept of finding and using an appropriate seed set for sentiment lexicon generation was further studied by Turney (2002), who developed a lexicon by observing word co-occurrences with the seed terms *poor* and *excellent* in a corpus derived from the once popular AltaVista search engine, whereby polarity was determined using the NEAR operator that returned relative sentiment-bearing terms within a fixed window.

The concept of using a set of seed terms to create a lexicon has been used in recent years as a corpus-driven approach to tune existing sentiment resources to new domains. For example, Mohammad et al. (2013) extended this method by observing tweets that contained a hashtag, a word in a tweet that denotes a particular topic or sentiment, that was based upon seeds derived from the synonyms listed for the words *positive* and *negative* in Roget's Thesaurus, to generate a Twitter sentiment lexicon for the classification of English tweets. Klebanov et al. (2013) worked on expanding a generic sentiment lexicon using a pivot-based paraphrasing and crowdsourcing method for use in classifying student essays that resulted in classification accuracy improvement of up to 15%, yielding an accuracy of 64.400%. The approach was also used to annotate the polarity of new words emerging in Chinese (Huang et al., 2014).

The corpus-based seeding method has also been recently used to develop multilingual sentiment lexicons for the sentiment analysis of Standard Arabic through use of machine translation techniques in conjunction with a number of Standard Arabic corpora (Eskander & Rambow, 2015; Mohammad et al., 2015), and to assign intensity scores to a sentiment lexicon (Sharma et al., 2015) by applying a similar technique to an intensity corpus (Pang & Lee, 2005) using seeds from a generic sentiment lexicon (Liu et al., 2005).

The dictionary-based approach to sentiment lexicon generation differs to the corpus-based approach by focusing on an existing lexical resource such as WordNet (Miller, 1995), and in doing so traverses different link-types in the given electronic dictionary in order to derive a

sentiment lexicon. One example is described by Esuli & Sebastiani (2005), who examine the definitions of words in WordNet to derive a sentiment lexicon. Similar to the corpus-based approaches, a seed set of positive and negative words is used to iterate over the lexical relations to generate the sentiment lexicon. This is expanded in further work (Esuli & Sebastiani, 2006) to produce the publicly available SentiWordNet lexicon, and further expanded again using a random walk of WordNet in later work (Baccianella et al., 2010) to produce refined sentiment scores in the SentiWordNet 3.0 sentiment lexicon. Kamps et al. (2004) take a similar approach to develop a sentiment lexicon whereby word polarity is determined by the link distance in WordNet from the seed words *good* and *bad*. Hu & Liu (2004) develop a similar approach to Esuli & Sebastiani (2005); however, as well as traversing the adjective synonym links of WordNet, antonym links are also traversed when constructing a domain independent sentiment lexicon.

Velikovich et al. (2010) developed a graph propagation framework to semi-automatically construct sentiment lexicons that combines the best of the dictionary-based and corpus-based approaches to lexicon construction. Instead of relying on the traversal of an existing lexical resource such as WordNet, in their work, a graph is built from co-occurrence statistics generated from 4 billion English web pages. This led to the formation of a resource that included a range of phrases including spelling variation, slang and multi-word expressions, that are not included in a number of previously derived sentiment lexicons.

Recently, this approach of combining the dictionary-based and corpus-based approach to lexicon generation (Velikovich et al., 2010) has been used to create sentiment lexicons for 136 major languages (Chen & Skiena, 2014). These resources were created using a graph-propagation technique that started with English sentiment lexicons (Esuli & Sebastiani, 2006; Liu et al., 2005) and through examination of a knowledge graph constructed from Wiktionary, the use of the Google Translate API, examination of transliterative links in closely related language pairs, and WordNet, the resources were able to be created. The more commonly spoken languages such as French and Polish yielded sizeable lexicons, containing 4,653 terms and 3,533 terms respectively, but languages that have not got a significant online presence in the

resources, such as *Quechua* or *Amharic*, only have minimalistic lexicons, of size 47 and 46 respectively. It remains to be seen whether these resources can produce adequate levels of classification performance in the given languages.

The final approach to sentiment lexicon generation is the manual approach. This approach can be seen as the most-time consuming approach to lexicon generation due to the logistical requirements involved in managing the annotators and ensuring the quality of the annotated resource, but for a specific domain, this approach can be seen as the most reliable and thorough method for lexicon construction (Schneider & Dragut, 2015). The first general purpose sentiment lexicon, the General Inquirer, was developed by Stone (1966) using the manual approach. This resource contains 11,788 words that have a positive or negative labelling, and this resource has been used as a basis for the examination of the coverage of newly generated sentiment lexicons (Hu et al., 2013a).

However, more recent work in the literature has shown that lexicons can be manually created in a way that is not overly expensive or time-consuming. Mohammad (2010) develop an emotion lexicon manually using a labour force consisting of five annotators on Amazon's Mechanical Turk platform[1], for less than $500, with a first run taking approximately 9 days to manually annotate and produce a lexicon of 14,182 terms.

Classification methods using a sentiment lexicon

Given a sentiment lexicon, an unsupervised approach to sentiment classification would use this as the basis of the classification. For example, Bollen et al. (2011) use the OpinionFinder lexicon (Wilson et al., 2005a) to determine the ratio of negative to positive terms in a tweet, and assign the greater ratio as the overall labelling of the tweet. OpinionFinder contains information regarding whether a positive or negative term weakly or strongly conveys a polarity, and collectively by selecting these parameters for the terms, a lexicon of 2718 positive words and 4912 negative words was used by Bollen et al. (2011) in the classification process. Hu et al. (2013a) define this as the *traditional lexicon-based method* whereby the presence of a positive

[1] https://www.mturk.com

word in a document receives a +1, and a negative word receives a -1, and the overall score is the summation of scores of all words in a document. Salah (2014) further define this as the averaging of counting of the sentiment-bearing terms identified by the lexicon.

O'Connor et al. (2010) also use the OpinionFinder resource, but they simplify the traditional lexicon-based method scoring as they are looking at tweet data, which they argue should be treated differently as the average document length they were dealing with was only 11 words, by labelling a tweet as positive if one or more positive words were present, and negative if one or more negative words was present. This resulted in some tweets being allowed to be both positive and negative, but as they were not considering classification accuracy but only the average sentiment of all tweets on a given day in relation to a topic, this appeared to work well for them.

The method of summing scores based on those assigned by a sentiment lexicon is extended by Choi & Cardie (2009) to examine and make use of the terms that explicitly alter the overall sentiment of a document, an algorithm that they name the Vote & Flip algorithm. In this procedure, words such as *not* are encoded in the lexicon with attached functionality that can affect the overall score of a phrase in a document if found. Using their method in tandem with the polarity lexicon developed by Wilson & Wiebe (2005) and the General Inquirer (Stone, 1966), this approach is shown to significantly improve classification accuracy ($p < 0.01$) on a test set of 400 documents from 64.2% accuracy when the method is not used, to 67.0% when it is applied.

2.2.2 Supervised approaches

The opposite of the unsupervised approach to sentiment classification is the supervised approach. Given a set of documents with associated sentiment category labels, the supervised approach to sentiment classification takes advantage of a subset of machine learning techniques to determine the likely set of rules or parameters that will form the basis of a learned classification function that can be used to automatically categorise unseen documents by the sentiment that they convey.

In this process, the labelled document set that forms the input is important in ensuring a robust model is learned. This does not totally eradicate human intervention however. Typically, human annotators will be required to label the document set appropriately prior to its input to the machine learning method, or at least confirm the robustness of any labels that may already be associated with a given dataset. This labelled document set is known as the gold-standard, and in learning a model, this is the label set that the model aims to achieve when classifying the documents. Increasingly, the labelling for the gold-standard is achieved in efficient and novel ways, such as webcrawling for data from relevant review sites (McAuley & Leskovec, 2013a).

In formalising the task of sentiment analysis as a supervised machine learning problem, we refer to the formal definition given by Manning et al. (2008) here for clarity.

As input, the learning process requires a document space $\mathbb{X}$, a set of training documents $\mathbb{D} = \{d_1, d_2, \ldots, d_n\}$, where n represents the number of training documents, and a set of classes $\mathbb{C} = \{c_1, c_2, \ldots, c_k\}$, where k represents the number of classes. The output of the algorithm for each d_i is a projected class $c \in \mathbb{C}$. The aim is then to learn an optimal classification function γ that accurately captures the mapping of the members of a document's set to classes for both $\mathbb{D}$ and future unseen documents:

$$\gamma = \mathbb{X} \rightarrow \mathbb{C} \tag{2.1}$$

The supervised learner that takes $\mathbb{D}$ as input and returns γ is referred to as Γ. Despite this formal difference, the learning method and the derived classifier are often referred to with the same name.

Given the learning of a classification function from $\mathbb{D}$, we are able to experiment with unseen data, which we will refer to as the test set. This can be used to evaluate the performance of γ. Evaluation metrics for γ will be detailed in relation to the relevant experiments that are run in this thesis, discussed section 4.4.1 of Chapter 4.

As well as the formalism of the learning process, we should also consider the representation of each document d. In order to map documents to the categories that represent them, each document should be represented in a computable manner. To enable this, each document in

the training set is transformed into a vector of features, whereby $d = (w_1, w_2, \ldots, w_n)$. Each dimension, denoted by w_i, encodes various document attributes such as word presence, word frequency, n-gram combinations, and so on.

Given this formalisation of a supervised machine learning classifier for sentiment classification, we will now give an overview of the supervised machine learning algorithms that have been applied to the task of sentiment classification, and discuss the performance of each classification model in turn.

Naïve Bayes for sentiment classification

Due to the intuitive motivation and speed of classification (Lewis, 1998), the Naïve Bayes (NB) classification model is one of the more frequently used models in the sentiment classification literature. When training, the NB classifier does not overfit the training data, meaning a reliable classification model should be generated given a suitable input (Ng & Jordan, 2001). However, the NB model is naïve in the sense that each feature for text classification is assumed to be independent of all other features. In assuming independence, the presence of a feature has no impact on the probability of another feature also being a member of the document vector. This is counter-intuitive to the workings of natural language, whereby meaning is derived from the company a word keeps (Firth, 1957). The independence assumption may seem flawed when considering its role in text classification, yet it has performed well in a number of classification tasks (Sebastiani, 2002).

In the sentiment analysis literature, Wiebe et al. (1999) were one of the first to examine the application of a variation of the NB classifier, the multivariate Bernoulli model, to the task of subjectivity classification. This model achieved an accuracy of 72.17%, and led the way for other researchers to apply the NB classifier to sentiment classification tasks. For example, Pang et al. (2002) demonstrated that when using a unigram feature set constructed from the words that appeared at least four times in a 1400 document corpus, an NB classifier was able to achieve a three-fold cross-validated accuracy of 78.7% when attempting to classify movie reviews.

More recently, Liu et al. (2014) examined the effects of modal verbs on the sentiment clas-

sification of product reviews. Results showed that the NB model outperformed an unsupervised lexicon-based approach and a support vector machine classifier trained for the same task.

Also, the classifier has been applied to the classification of sentiment in tweets (Talbot et al., 2015; Gamallo & García, 2014; Dermouche et al., 2013), where it has been found to outperform other supervised machine learning methods, including again, the support vector machine classification model.

Support vector machines for sentiment classification

The support vector machine classification (SVM) model was developed by Vapnik (1995) and gained popularity in the text classification literature due to the work of Joachims (1998). It is based on the principle of structural risk minimization (Vapnik, 1995). This concept aims to find a hypothesis that will yield the lowest possible error when classifying a given data set. SVMs implement this idea by finding the hypothesis that minimizes the boundaries of true error, and in the process, learns a linear threshold function. It does so through the use of a kernel function, that maps non-linear data into linear space in order to create a decision boundary (Joachims, 2002).

In the sentiment analysis literature, Pang & Lee (2005) experiment with three permutations of the basic SVM for sentiment rating prediction. They experiment with the *One-vs-all* approach (Rifkin & Klautau, 2004), a methodology used for multiclass SVM classification. This is formulated as n separate binary classifiers, each competing to classify the test data with the most appropriate labelling given the greatest signed distance from the generated decision plane. Second, they experiment with a support vector regression approach, also used by Wilson et al. (2004) to detect strong and weak opinion clauses. Finally, they experiment with a metric-labelling approach, that attempts to use instance and label similarity to penalise the initial classifier if it attempts to classify similar instances with differing labels. Accuracies of the given setups range from approximately 36% to 75% over four different datasets, and finds that the one-vs-all approach was significantly weaker than the others approaches when classifying review ratings.

Whitelaw et al. (2005) investigated the use of an SVM classifier to automatically categorise the movie review dataset developed by Pang et al. (2002). Instead of the typical positive and negative document distinction that a sentiment classifier is trained upon, in their work, Whitelaw et al. (ibid.) investigate the use of a taxonomy of appraisal group features that are based on Appraisal Theory (Martin & White, 2005). The choice of the SVM classifier appears to be somewhat of an arbitrary decision, however results confirm that an accuracy of 90.2% is able to be achieved when the classifier is trained on a lexicon of appraisal group features. These consist of attribute values over a set of appraisal group taxonomies.

SVMs have been used for a number of subtasks in the sentiment classification literature. They have been implemented for cross-language sentiment classification (Wan, 2009) to good effect, and sarcasm detection in the domain of sentiment analysis of tweets (González-Ibáñez et al., 2011). Jindal & Liu (2006) implement an SVM classifier to categorise comparative sentences into four types, which is shown to perform competitively in comparison to an NB classifier.

State-of-the-art results have been achieved using an SVM classifier trained with a linear kernel for the task of short message classification (Mohammad et al., 2013) on the SemEval 2013 *Sentiment Analysis in Twitter* task (Nakov et al., 2013). Using this classification model, their approach achieved an F_1 of 0.889 on a tweet phrase contextual polarity disambiguation subtask, and F_1 of 0.690 on a tweet classification subtask. Other works that focused on the classification of tweets in the literature were also able to achieve competitive results (Karanasou et al., 2015; Jaggi et al., 2014).

Logistic regression for sentiment classification

Logistic regression (LR) is a discriminative classification model that determines the most distinctive features for classifying a text (Ng & Jordan, 2001). It is sometimes referred to as maximum entropy modelling (Manning et al., 2008).

In the sentiment classification of social media comments, it performs comparably to an SVM in classification experiments, achieving 58.5% accuracy, and significantly outperforms

an NB classifier (Thelwall et al., 2010). In a multi-domain customer review classification task, Xia et al. (2011) also find that the SVM and LR classifiers perform comparably, each achieving approximately 85% accuracy over a number of experiments. This comparative behaviour between SVM and LR was similarly discussed by Wang & Manning (2012). On review data from restaurants, laptops and hotels, Hamdan et al. (2015) develop an LR model that achieves a maximum accuracy of 85.54% when classifying data from the hotel domain.

LR models have also been applied to the sentiment classification of tweets. Alhessi & Wicentowski (2015) attempt the ternary sentiment classification task (Rosenthal et al., 2015) using a logistic regression model trained with a one-vs-all configuration. They achieve an above average F_1 of 0.584 on the task, but they do not outperform the task winner (Hagen et al., 2015), who achieved an F_1 of 0.648.

Random Forests for sentiment classification

The random forest (RF) classifier is an ensemble classification method that combines the prediction of a series of decision trees trained on different subsets of the training data, known as bagging, in order to carry out classification (Breiman, 2001).

In the sentiment classification literature, it is relatively underused in comparison to a technique such as support vector classification, despite its ability to overcome the overfitting effects of that some classifiers may exhibit. It has been found to outperform other supervised models, such as decision tree and logistic regression, in the classification of both sentiment in social media posts (Zhang et al., 2011) and online forum messages (Ofek et al., 2013). In a comparison with some commercial sentiment classification tools, such as Alchemy[1] and Lymbix[2], an RF classifier yields significant improvements in classification performance, with a 9% increase in accuracy over the commercially available tools (Cieliebak et al., 2014). Using the random forest method on the SemEval Twitter sentiment classification task (Rosenthal et al., 2015), Uzdilli et al. (2015) achieve an F_1 of 0.626.

[1] `http://www.alchemyapi.com`
[2] `http://www.lymbix.com`

Deep learning for sentiment classification

LeCun et al. (2015) define deep learning as the process of developing computational models that consist of a number of processing layers, each of which is able to learn multiple abstract representations of a data set. With each level of abstraction, the deep learning process is able to model intricate structures and determine the important aspects of the input data that are discriminative when used in a classification task, such as sentiment analysis.

In particular, the convolutional neural network or ConvNet (LeCun et al., 1998) has been adapted and applied to the problem of sentiment classification. This neural network model applies convolving features to each layer of the network, a process that has proven useful in other areas of natural language processing, such as semantic parsing for question answering (Yih et al., 2014).

Using this technique, Kim (2014) trains a convolutional neural network for a number of sentence-level classification tasks, including the binary sentiment classification of sentences in the movie review dataset (Pang & Lee, 2005) and the customer review dataset (Hu & Liu, 2004). On both datasets, the application of the convolutional neural network was able to produce state-of-the-art accuracy results of 81.5% and 85.0%, respectively. This method improves on previously applied deep learning techniques to the task such as the matrix-vector recursive neural network technique (Socher et al., 2012), that achieved an accuracy of 79.0% on the movie review classification task, and Zhou et al. (2014), who develop a semi-supervised hybrid deep-belief network consisting of a number of hidden layers that are constructed using restricted Boltzmann machines, who achieve an accuracy of 72.2% on the movie dataset. Convolutional neural network have also achieved an accuracy of 86.4% when attempting to classify the sentiment of Twitter data (dos Santos & Gatti, 2014).

Poria et al. (2015) further examine the application of a deep convolutional neural network trained on a corpus for multi-modal sentiment analysis of short video clips. Combining textual, visual and audio features, they achieve an accuracy of 86.27% on a dataset of 447 short videos (Morency et al., 2011), in which a person talking utters a single sentence that either conveys a positive or negative sentiment.

While reporting strong performances on a number of classification tasks, deep learning can potentially be a computationally complex approach to classification, with model development and application potentially spanning weeks (Ciresan et al., 2010). This is not desirable, especially when other modes of classification that perform comparably, such as the previously discussed approaches, are able to perform training and evaluation significantly faster. Due to this, the work in this thesis will focus on the aforementioned machine learning models.

Ensemble methods for sentiment classification

Ensemble learning as an approach to classification combines several classifiers in order to form a potentially more discriminative classification model (Rokach, 2010). This method has gained some traction in the sentiment classification literature due to its ability to correctly classify in unison what individual classification models may potentially misclassify on their own. However, it is not as widely investigated as the models that we have previously discussed in this section, as a single model can still be a complex entity to decipher when attempting to analyse any misclassifications that it produces.

Ensemble learners differ from each other based upon not only the selection of classification models, the base learners, that are combined in an ensemble, but also the method of combination that is applied to them. Common approaches to combination include bagging (Papakonstantinou et al., 2014), boosting (Dubout & Fleuret, 2014) and stacking (Li et al., 2015). Bagging trains a number of base learners on bootstrapped training data and applies a voting protocol to determine the overall classification outcome (Breiman, 1996). Boosting follows a similar concept to bagging; however, the base learners are trained upon weighted versions of the training set, whereby the weightings are dependent on the base learner's past individual performance on the classification task (Zhang et al., 2014). The last method, stacking, aims to reduce the error rate in classification through a process that splits the training data, then trains several base learners on the first part, and tests these base learners on the second part, and finally uses the predictions as the inputs to train a further, potentially more discriminative model (Wolpert, 1992).

Using a bagging technique, Andreevskaia & Bergler (2008) develop an ensemble classification system that combines the output of an unsupervised lexicon-based system with an SVM classifier trained on in-domain data using a weighted voting technique. Examining the performance of this system shows classification improvements over the application of the classification techniques individually over four separate domains, leading to a maximum accuracy of 78.0% and F_1 of 0.88 on the product review domain (Hu & Liu, 2004).

Ensemble methods have also been examined for state-of-the-art natural language processing challenges. For example, Hagen et al. (2015) develop a meta-classifier for the SemEval 2015 Twitter classification task (Rosenthal et al., 2015) that is an ensemble of the previous years best classifiers on a similar sentiment classification task in the same domain. By taking the four previous best-performing systems, three from 2013 and one from 2014, each was reimplemented to yield a confidence score for classifying an instance into a particular category, either positive, negative or neutral. In the work of Hagen et al. (ibid.) the confidence scores for a given class from each of the classifiers is averaged to determine the relevant classification, yielding state-of-the-art results on the 2015 task, with an F_1 of 0.648. Wicentowski (2015) also apply an ensemble of 23 classifiers to the same task, that were mainly different variations of the standard machine learning classifiers discussed previously in this section. Using a weighted voting protocol on the combination of the classifier's outputs, achieved an F_1 of 0.619 on the sentiment classification task.

2.2.3 Challenges

At training time, the data used by supervised machine learning approaches to sentiment classification can introduce a number of dependencies to the learned model that may prove challenging when categorising documents by the sentiment conveyed, and could have the ability to negatively affect a system's classification performance.

The first of these types of dependencies, explored by Engström (2004), is topic dependency. In her work, she draws focus to the fact that training and testing data for sentiment classification typically discusses a single topic, such as film reviews, which makes it difficult to verify

whether classification features are related to general sentiment conveyance in text, and not just the topic of the documents being classified. In experiments on news articles of varying topicality, financial documents, and mixed articles from various subject domains, the assumption of topic dependence in sentiment classification using the linear SVM classifier was confirmed when drops in accuracy were found when training and testing across the three datasets. To overcome the problem of topic dependency in sentiment classification, a hand-coded list of sentiment words was investigated, but found to only introduce a data-sparsity issue when used in combination with the SVM classification model, and results using this approach were still worse in across topic classification than within.

The second of the dependencies formed by sentiment classifiers posed with the challenge of classifying data from varying sources is the broader problem of domain dependency. This expands upon the issue of a sentiment classifier being biased towards the topic that it is trained upon by examining the effects of different domains when classifying sentiment across data from distinct document sets from different sources, such as product reviews to news articles. Aue & Gamon (2005) examine this problem with respect to four domains: film reviews, book reviews, product support service data and knowledge base data. The work establishes that cross-domain classification on the aforementioned domains is generally poor, and emphasizes the fact that domain differences in sentiment conveyance are substantial to the point where a classifier trained on one domain is barely able to produce results that surpass the arbitrary baseline in another domain. The work notes that the difficulty of sentiment classification in different domains varies widely, with film reviews achieving the best accuracy using an SVM classifier trained on the top 20,000 n-gram features calculated using the log-likelihood ratio (90.45%), and the knowledge base web survey data yielding an accuracy of 77.34%. In attempting to solve the issue, the best approach was found to combine data from the target domain into the classification process using an NB classifier trained using the expectation maximization algorithm, which performed significantly better than merely combining all the training data from all the domains, or using an ensemble of classifiers for the task.

Since the work of Aue & Gamon (2005), a number of articles in the literature have tack-

led the issued of domain dependency in sentiment classification with respect to a benchmark Amazon dataset developed by Blitzer et al. (2007). This dataset consists of reviews from four domains: books, electronics, kitchen and DVDs. Work on this dataset to minimize the transfer loss across domains have investigated the use of methods including spectral feature alignment (Pan & Yang, 2010), a sentiment thesaurus that is sensitive to multi-domain sentiment (Bollegala et al., 2011), a join sentiment topic model (He et al., 2011), and a deep learning approach (Glorot et al., 2011). Each has had increasing degrees of success in minimizing the transfer loss that occurs in cross-domain classification, and the last paper was able to yield marginal increases on the in-domain transfer for some of the domains when tested on the benchmark dataset.

The third dependency that classifiers trained for sentiment analysis may be subject to is a temporal dependence. This dependence is likely to be associated with classifiers trained on social media data whereby the sentiment in regards to a topic has the potential to change over time. Read (2005) examines the phenomenon when classifying film reviews in the Polarity 1.0 dataset (Pang et al., 2002). A classifier is first trained and tested on reviews taken from the years leading up to 2002 , and then trained again on film reviews with the same distribution collected from the years 2003 and 2004. A temporal dependence was confirmed when drops in classification accuracy were recorded across the two datasets. Although this appears to suggest temporal dependency, it is not clear whether this drop could be attributed to other aspects, such as differences in author style amidst the training data.

The fourth dependency that sentiment classifiers may be sensitive to is genre dependency. Genre is very much an influential factor on the written style of a document, and so it follows that training and testing a sentiment classifier across documents of different genres should lead to negative effects in the outcome of classification due to this dependency. Mejova & Srinivasan (2012) examine the effects of genre while controlling for the effects of variable topic in blogs, tweets and reviews. Results found in experiments using a logistic regression classifier with ngram features, whereby n was limited to three. Reviews were found to generalise well to classify sentiment in Twitter and in blogs, with tweets following as a source of training data

that was able to generalise well. While other researchers found that the combination of data from mixed source across domain didn't yield improvements in classification (Aue & Gamon, 2005), it was found to do so in the cross-genre experiments.

These dependencies are problematic for sentiment classification going forward, as it either forces the scope of sentiment classification to be very narrow, or requires data intensive approaches to enable classifiers to suitably generalize. One specific dependency that is not discussed in the literature, but would undoubtedly affect the outcome of sentiment classification, is a document set that contains documents with different purposes, what we shall refer to as the document or review *type*, that a classifier is trained and tested upon. In such a case, a set of documents in the review genre may belong to a single domain, such as cars, discuss a single topic, such as steering wheels, at a single time point in time, but the review type that is used to train a learned model has the potential to affect classifier performance. For example, one review may list a reviewer's likes and dislikes, while another review may instead be attempting to give advice. Both have the potential to convey sentiment within the review, but we hypothesise that sentiment would be communicated differently across the different types of review. This has not been explicitly examined in the literature and so this thesis will examine this problem further when considering the task of sentiment classification in the clinical domain.

2.3 Sentiment Analysis in the Clinical Domain

Thus far, this chapter has given an overview of the literature on general approaches to sentiment classification. In this section, the work focusing on the research and development of sentiment analysis in the clinical domain will be discussed. This work can largely be examined by the type of document that sentiment analysis methods are developed to classify: biomedical texts or instances of patient feedback. An examination of the work that focuses on these two types of document demonstrates that the traditional definition of sentiment requires expansion and clarification to understand what it means for a clinical document to be positive or negative, which is discussed in the final subsection.

2.3.1 Analysis of biomedical texts

Biomedical texts detail the interaction between a patient's health issues and the methods used to treat them. A biomedical text describes whether a treatment was successful or unsuccessful at dealing with an illness, and how this treatment has affected the state of the patient. These documents can therefore be subject to automatic analysis with sentiment classification techniques to generate summaries of the relative polarities that a biomedical document conveys about a particular drug or patient type. Documents for evaluation focus on the use of article abstracts scraped from PubMed[1] or Clinical Evidence[2], or specialised medical document repositories such as MIMIC II[3].

Biomedical texts are used by evidence-based medicine practitioners to ensure the best-possible decisions can be made based that upon the outcome of clinical research. However, the large number of biomedical texts often makes the task overwhelming. This can be alleviated through the use of sentiment analysis tools to enable the generation of a polarity based summary regarding the usage of a drug or treatment from a set of clinical outcomes in the biomedical texts. For example, Niu et al. (2005) developed a system to classify the polarity of clinical outcomes in medical texts that attempted to answer clinicians questions detailing how effective a particular treatment was. In their work, a clinical outcome could be positive, negative, neutral or simply yield no outcome. An SVM classifier was trained to classify a set of documents from the Clinical Evidence website into one of the four classes. Using a combination of unigram, bigram, change phrases, negators and category features, the best classification accuracy of 79.4% was achieved. Of the features used, change phrases relating to an increasing or decreasing fluctuation in clinical values were found to be useful indicators of sentiment that boosted classifier performance. Despite this relative boost, experiments were not cross-validated, so an uneven class distribution could have caused a skew in the resulting classifier performance. However, Sarker et al. (2011) also use a SVM classifier trained with a similar feature set on a different dataset trawled from PubMed and the Journal of Family Practice to

[1] http://www.ncbi.nlm.nih.gov/pubmed/
[2] http://clinicalevidence.bmj.com/
[3] http://www.physionet.org/mimic2/

achieve a similar accuracy of 74.9%. Despite working on a smaller dataset, the neutral category is not considered for manual annotation in their work. Results of manual annotation between four annotators produced a Fleiss' Kappa of 0.706, indicating that there is agreement beyond chance between the mutual documents the annotators labelled and that the categories were well defined.

In a similar vein to the outcome-focussed classification of medical abstracts, Deng et al. (2014) classify the sentiment in the clinical narratives given by physicians and nurses. These documents express the professional opinions and judgements about the health of a patient as opposed to the academic nature of published biomedical documents that were the basis of previous work. Again, sentiment analysis is not undertaken in the traditional sense and in this work it refers to the information on the health status of a patient or the seriousness of a symptom, where polarity is relative to status changes such as an improvement in a patient's health. The texts are first POS tagged and compared to a subjectivity and sentiment lexicon. A linguistic analysis in this work shows that traditional sentiment analysis resources do not provide total coverage for of adjectives and adverbs that are used in biomedical documents. Owing to this, the results of sentiment analysis are fairly poor by use of traditional lexicons alone. Deng et al. (2014) conclude that more sophisticated methods may be required in order to handle anomalies such as typos in nurses letters, and context for phrases such as *blood pressure decreased* when using lexicon based approaches.

Suicide notes have also been the subject of a biomedical sentiment classification task to determine suicide susceptibility (Pestian et al., 2012). Suicide notes were classified into a choice of sixteen categories describing the emotions present in the notes. Multiple approaches were applied, and the best F_1 achieved was 0.6139 by (Yang et al., 2012) using a voting protocol alongside an ensemble classification method combining SVM, NB and Maximum Entropy classifiers.

2.3.2 Analysis of patient feedback

As well as its application in evidence-based medicine, sentiment analysis has also been applied to the discussions, views and concerns that patients leave in various forms of feedback. Unlike biomedical texts, the content of patient feedback does not necessarily focus on the success rate of a treatment, but rather on the more general patient experience, given from the patient's perspective. Hence, patient feedback can be likened to more traditional data sources for sentiment analysis, such as product reviews. The source of patient feedback data tends to be specialist patient feedback portals, such as PatientOpinion[1] and RateMDs.com[2] or medical forums focussing on particular medical issues such as IVF[3].

The suitability of online patient feedback data is studied by Hopper & Uriyo (2015). In their work, a corpus of reviews of gynaecologists from RateMD.com is examined. This was an exploratory study into the applicability of automatic sentiment analysis, in particular, observing time between complaints to predict when the next complaint will occur. They found that there was a 97.97% agreement between the labels that RateMD.com users gave to their own comments and separate, blind labellings given to the same set of comments by the study's authors. They therefore conclude that the use of such data for sentiment classification is complimentary to the assumptions of a group of manual annotators, and therefore is suitable for the development of models for classifying reviews of gynaecologists in particular.

When considering the methods applied to classify the sentiment of patient feedback, traditional methods from text classification and sentiment analysis have been applied. For example, the lexicon-based approach to the sentiment analysis of patient feedback is a widely discussed technique in the literature. Goeuriot et al. (2012) develop such a lexicon for sentiment analysis in the clinical domain. They first merge SentiWordNet 3.0 (Baccianella et al., 2010) and a precompiled subjectivity lexicon developed by Wilson & Wiebe (2005). Following this, using a corpus-based approach for keyword extraction, they augment the merged lexicon with salient words from reviews in a drug expert forum. This procedure produced 1446 additional terms,

[1] https://www.patientopinion.org.uk
[2] https://www.ratemds.com/
[3] http://www.ivf.ca/forums

of which 1142 are not in the merged general sentiment lexicon. The lexicon forms the basis for a vote-flip sentiment labelling process. This process counts all positive and negative words present in a review to find the predominant polarity and given the presence of negation words then flips the predominant sentiment. However, application of this algorithm to a test set of 25,000 comments leads to a low accuracy, less than 50%, which they in part attribute to the presence of neutral comments, those that do not explicitly convey a sentiment, in the test set. It could be argued, however, that the original sentiment lexicons are more suited to processing review sentiment as opposed to comments made by patients about healthcare services, whereby the in-domain terminology used to express sentiment may differ slightly to its expected traditional usage.

Another lexicon-based approach is developed by Sokolova & Bobicev (2013) to examine the sentiment of posts on an IVF medical forum. They argue that the traditional positive-negative categorisation is not sufficient for the classification of medical posts and therefore propose the categories of encouragement, gratitude, confusion, facts and facts that convey encouragement for classification. In their study, they annotated a set of medical forum posts on infertility. Of the 752 texts annotated, 150 were annotated as uncertain. They found that typical sentiment lexicons, SWN and WNA, did not provide significant coverage and so they developed a lexicon, HealthAffect. This was developed using an adapted version of the Pointwise Mutual Information Turney (2002) algorithm that considers unigrams, bigrams and trigrams where the frequency was greater than five in the medical texts as candidates for use in the final lexicon. Using an NB classifier in conjunction with HealthAffect, they achieve a precision of 0.527, a recall of 0.541 and an F_1 of 0.518 for the six-class classification task. This proves to be significantly better than using WordNetAffect alone (precision = 0.322, recall = 0.350 andF_1 = 0.303), and the choice of NB classifier was found to be significantly better than the k-nearest neighbours classification approach, which yielded a precision of 0.377, recall of 0.376 and an F_1 of 0.340 when using the HealthAffect lexicon.

Machine learning based approaches have also been developed to classify patient feedback by the sentiment expressed. Xia et al. (2009) develop a multi-step classification approach to

opinion mining in patient feedback. Their approach is based on the hypothesis that sentiment and topic are related. They give the example that where a patient comments about the parking of a hospital, these comments are going to be negative. They use a dataset of 1200 items of patient feedback from Patient Opinion. Due to the assumption that topic and sentiment are related, by first applying a topic classifier and then a sentiment classifier to a test set, they observe an F_1 score of 0.77 in comparison to 0.656 for the single step approach.

Cambria et al. (2012) also observe data from Patient Opinion in their work. They develop combined emotion and polarity categorisation framework they refer to as sentic categorisation to capture patient-reported outcome measures (PROMs). Using a variety of lexicon and unsupervised clustering based techniques, they claim to be able to exploit the semantics of patient opinions to aggregate and evaluate a patient's health status. Evaluation claims to yield 91.0% accuracy on a dataset of 2000 posts from PatientOpinion, although other metrics are not given. When testing the system on LiveJournal blog posts, detection of posts with a *happy* mood yield F_1 score of 0.82, and *sad* 0.74. It is not noted whether these posts were from the clinical domain.

Georgiou et al. (2015) examine the difficulty of analysing sentiment in the healthcare domain, and find the use of the NB algorithm to be better than the commercially available sentiment analysis platforms, Semantria[1] and TheySay[2]. They report an average accuracy of 82.4% using four-fold cross-validation, however the dataset only contains 137 documents, and the data is highly skewed towards the negative category. The ability to detect sarcasm was also examined, and again, they found that the NB classifier was found to yield superior performance to commercially available software that claimed to be able to do this task.

Greaves et al. (2013) investigate the application of supervised machine learning classifiers to categorise sentiment in a subset of the NHS Choices data. They examined three areas on a binary scale: if a patient thought a hospital was clean, if a patient thought they were treated with dignity, and if the patient would recommend the hospital, based upon the unstructured patient comment[3].

[1] `https://www.lexalytics.com/semantria`

[2] `http://www.theysay.io`

[3] These qualities were not asked for on the website after 2010, and were therefore not available in the data we used.

The work of Greaves et al. (ibid.) gives a comparison between the results of traditional paper-based survey results and the accuracy of the machine learning classifiers trained for sentiment analysis. The comparison examined three aspects of the surveys: whether the patient thought the hospital was clean or dirty, whether they were treated with dignity or not and whether they would recommend the hospital. Results of the machine learning classifiers tested were compared to the paper based survey results using the Kappa statistic and the Spearman correlation coefficient. The best Kappa statistic of inter-rater reliability was found to be between 0.4 and 0.74, depending on the quality compared when using an MNB classifier. Also when using this approach, the Spearman correlation coefficients were between 0.37 and 0.51, highlighting a weak to moderate association between the machine learning model's predictions and the outcome of the paper based survey.

The results of the machine learning experiments carried out by Greaves et al. (ibid.) demonstrate that the use of the MNB classifier yielded the best performance when classifying the overall rating of a patient review. When using the MNB model for classification, the accuracy was 88.6% and F_1 was 0.89, with the results for the decision tree, bagging and SVM approaches ranking behind this method. Similar classification tasks into the classes of cleanliness and dignity also followed this results trend.

The work is able to make use of the implicit notion that if a patient says they would recommend a hospital, they are then implying a positive sentiment, and conversely if not a negative sentiment, therefore automatically extracting a nominal categorisation. This is similar to a number of approaches that automatically scrape reviews with rating information from web pages. Greaves et al. (ibid.) note that this eliminates a responder bias being formed in regards to the comment. However, in this thesis, we argue that where no quantitative rating is left alongside a comment, we must look to other sources of information for rating induction.

One weakness of the approach Greaves et al note is that without context it is difficult to classify phrases such as 'cup of tea'. This is not commonly used in many domains, and therefore does not incite a sentiment-bearing reaction. A system developed for general sentiment classification would not be able to handle this. Context must be sought from relevant, yet external

information to the utterance that is being classified. This thesis proposes a potential solution to incorporating such information into the classification process.

Sentiment analysis has also been applied to patient review data from other sources. Sharif et al. (2014) develop a rich feature framework, that is tested on both forum posts from the Ask a Patient website and a set of pharmaceutical tweets that they refer to as Pharma. The framework that they develop is feature rich in the sense that it uses a wide range of sources to construct the feature representations used to train an SVM ensemble, including baseline features such as n-grams, semantic features based upon WordNet mappings, emotion-related features from SentiWordNet and AffectWordNet, and finally domain specific features that focus on medical entities in a text. In combination, on the Ask a Patient data, an accuracy of 78.2% is achieved, and 79.73% accuracy is achieved on the Pharma data. The sentiment of public health related tweets is also studied by Bobicev et al. (2012), and results mimic that of the Pharma data study, achieving an accuracy of 82.4% using an NB classifier with features that are consistent with the class labels of the training data. Ali et al. (2013) also examined the classification of medical forum posts using an NB, SVM and LR classifier, however the best-performing model, the LR, only achieved an F_1 of 0.685, with a precision of 0.688 and a recall of 0.686, when using lemmatization features.

The analysis of communities of patients and the feedback that they give in online forums has also been studied. In particular, sentiment analysis has been applied to classify posts and determine changes in thread sentiment in an online community of cancer survivors (Qiu et al., 2011; Biyani et al., 2013; Ofek et al., 2013; Portier et al., 2013).

Beyond the classification of sentiment in patient feedback, the classification of emotion in patient forum posts has also been analysed by Melzi et al. (2014). With more categories to classify the posts into, this is naturally a more difficult task, and Melzi et al. (ibid.) cite the lack of consensus between the sixty annotators who took part in the initial labelling of the data as a difficult aspect. This trend followed in machine learning experiments that were ran using an SVM classifier: the best F_1 for the multiclass experiments was 0.258 and for a positive-negative classification experiment 0.657.

Another aspect of analysis amongst communities of patients is the degree of influence that they have on one another (Zhao et al., 2014). A metric that captures the degree of influence is developed that observes the nature of sentiment-bearing interactions between patients, and if the sentiment of the original poster is determined to change, then a high degree of influence is applied to the users who replied. Results of experimentation highlight that a higher positive sentiment in the reply is found to be indicative of a greater positive change in the originator's sentiment.

2.3.3 Expression of sentiment in the clinical domain

The concept of sentiment is a loosely defined notion that centres upon the expression of polarity in a text. When considering how a text may convey a sentiment in the clinical domain however, as we have shown in the previous two sections, what constitutes a positive or negative sentiment expression in text varies dependent on the type of text that is being considered in the clinical domain and therefore needs elaboration and refinement. One such example is the expression of sentiment in a clinical document discussing drug usage effects, which is not typically encoded into a generic knowledge base and therefore is not as simple to computationally determine the sentiment of a document. Denecke & Deng (2015) capture the following characteristics of sentiment that may be exhibited in documents from the clinical domain:

- The change in the health status of a patient e.g., *improving, worsening*.
- The effect of medical conditions on a patient e.g., *benign tumour*
- The certainty of a patient's medical condition or ability to treat e.g., *unsure*.
- The outcome of medical treatment on a patient e.g., *removed the cyst*.
- Opinions on aspects of interaction with the health service e.g., *caring nurse*.

These five aspects exhibit elements that are somewhat unique to the clinical domain. It is unlikely that a general purpose lexicon will incorporate specific clinical terminology such as

benign tumour, or label this phrase correctly. For this reason, the remainder of the thesis will focus on the ability for supervised machine learning methods to model the sentiment conveyed by patient feedback, in respect of different types of patient review, and apply the developed methods to unseen documents to verify their robustness.

Summary

In this chapter, a review of the relevant literature in the field of sentiment analysis has been undertaken, and state-of-the-art approaches to the problem have been discussed. While a number of solutions partially solved the problem, currently proposed techniques for sentiment classification are found to have weaknesses, and the review of the literature in the field culminates with a discussion of the problematic dependencies that sentiment classifiers learn in training. Such problematic dependencies include the issue of review type dependency, which is an issue that to our knowledge has not been examined in the literature, and so, this will be further examined over the course of this thesis, with suitable methods being proposed to overcome this difficulty.

The second section of the literature review examines the work on the development of sentiment classification techniques to classify texts in the clinical domain. In examining and discussing this work, we find that sentiment analysis has been used to summarise and analyse sentiment in biomedical texts, but primarily the literature has focused on the analysis of patient feedback from sources such as social media, forums and health provider pages. Supervised machine learning methods have been used in the literature to classify such data, and this family of techniques yields a higher standard of classification than using unsupervised approaches to the problem. However, no consensus can be found over a preferred supervised machine learning classification method to use for sentiment classification in the patient feedback domain. Therefore, through extensive experimentation, this will be explored further in Chapters 4 and 6.

CHAPTER 3

DATA ANNOTATION AND ANALYSIS

Introduction

This chapter will first describe the nature of the data that we are examining in this thesis. Patient feedback data is little studied in the field of sentiment analysis, and so in this chapter, the domain will be described, sources for patient feedback will be identified, and the structure of patient feedback will be defined, along with the role of organisational responses to the feedback.

We will then describe the annotation process of the NCSD. Unlike the Type 1 reviews, the Type 2 reviews do not have an explicitly labelled sentiment. Therefore, the first phase of the annotation is carried out on a subset of the Type 2 reviews from the NCSD. As we will discuss, the organisational responses to the reviews may be indicative of the overall sentiment of a patient feedback instance. To test this assumption, the responses to the subset of the Type 2 reviews that are manually labelled are also subject to a rigorous annotation process for the sentiment that the organisation response is replying to. A comparison between the annotated feedback and response documents reveals a substantial degree of agreement between Type 2 review sentiment and the sentiment that the related response appears to be replying to, which gives us confidence in investigating the use of responses in automatically classifying a review's sentiment, a topic that we will discuss in future chapters of this thesis.

Following the discussion of the annotation process, a corpus analysis is carried out on the reviews and responses in the NCSD that examines the linguistic similarities and differences

between Type 1 and Type 2 reviews, the content and structure of positive and negative patient feedback, and the language that is used by an organisation when responding to an item of patient feedback. A frequency analysis highlights similarities in reviews of differing sentiments from Type 2 reviews, indicating that similar topics are being discussed but from different points of view. A part-of-speech analysis is incorporated into the corpus analysis that highlights the higher frequency of adjective and proper noun use in positive reviews compared to negative reviews, but also the more frequent use of verbs and adverbs in negative reviews. Finally, the most representative terms of the patient feedback are examined in a keyness analysis. Three stages are carried out in the keyness analysis: first, a comparison to the British National Corpus (BNC), followed by a keyness analysis between review types, and finishing with a keyness analysis between positive and negative items of patient feedback. Results of the keyness analysis follow the trends of the part-of-speech analysis, and reveal that adjectives are commonly used markers of positive sentiment in patient feedback, whereas the markers of negative sentiment are somewhat more subtle. A corpus analysis of the responses finds that they are useful indicators of comment sentiment, and within them, there are commonly used, stereotypical aspects that are explicit markers of original comment sentiment that could be useful if factored into the sentiment classification process of patient feedback.

It should be noted that throughout this chapter, highly frequent or keywords discussed in each section will be italicised.

3.1 Patient feedback data

The rapid development of the internet as a medium for interaction has enabled its users to share their judgements, beliefs and experiences through easy to use, socially curated websites (Lee & Ma, 2012). This has generated vast amounts of opinion-bearing review data that is of interest to large companies and market researchers (Duan et al., 2008). This data is typically unstructured and requires relevant natural language processing techniques to give interpretable structure to the data. In the sentiment classification literature, the focus has been on customer reviews as

these discuss the positives and negatives of company specific products and enable actionable insights to be discovered through a process of opinion mining (Pang & Lee, 2008). While customer reviews enable companies to tweak their product lines and respond to criticisms, thereby protecting their margins, there are far more essential reviews online that could be used to improve a person's standard of life; namely patient feedback (Trigg, 2011).

In this section, we give an overview of the aspects of patient feedback data relevant to this thesis. We discuss the information available in patient feedback, how the feedback may be structured, and the online sources where patients are able to share their views and browse the experiences of others.

3.1.1 Domain description

The Picker Institute (2015) gives a concise description of patient feedback:

> *"Patient feedback consists of the views and opinions of patients and service users on the care they have experienced".*

In the field of sentiment analysis, this is a relatively understudied domain when compared to the domains of product or film reviews. However, as data has gradually become more open and accessible in similar ways to film and product reviews, this domain has shown the potential for its use in developing sentiment classification techniques.

This domain is quite unlike other data studied in sentiment analysis in the way that people view and interact with it. For example, if a film review is bad, the chances are that someone will not see that film. Similarly, if reviews are positive, this will positively affect the number of people seeing the film, and the box-office profits. These effects can also be produced in patient feedback, but having an illness or injury is unlike the interest to watch a film, and will, therefore, be treated in a different, somewhat more serious manner.

For these reasons, the feedback will undoubtedly be rich in opinion, which makes the domain enticing for the development of sentiment classification algorithms. As previous work has shown, this is true, and hence modelling such data is a challenge for current techniques.

The domain of patient feedback is fixed in its nature, again making it attractive data for sentiment classification. Classifiers that are trained do not need to worry so much about training on data that is heterogeneous, and therefore too shallow, thereby generating weak models. The overall domain of patient feedback is the healthcare system, and the content of a review will tend to focus on a small collection of aspects associated with this. Given a patient reviewing his GP practice, the number of aspects that can be discussed in a review is limited to the uniform process that is undertaken when attending a surgery. The only individual aspects will come from the subjective interactions of the patient during that time. This would not be the case in blog posts or tweets, whereby a multitude of different topics could be the subject of the document. Additionally, the metadata associated with the review, such as time and place, anchor the review to a particular real life entity. This, in turn, makes the feedback actionable: enabling a feasible decision process to be taken post review by those within the health service reviewing the review, or by someone browsing the reviews attempting to find the best place to seek medical treatment.

In addition to the feedback being actionable, a discourse can be generated given an initial review. Given the feedback, a management representative can respond to the feedback, either clarifying or seeking clarification on aspects mentioned in the initial review. This response mechanism seems to be unique to the online feedback where the company is providing the services, such as Trip Advisor, however, it is not always incorporated into the sentiment analysis process. As this thesis will explore, the response can be used to gauge the sentiment of the review or to clarify ambiguous instances of opinion conveyance, where a model may not have detected the presence of polarity in a document.

Just as the domain is fixed, the perspective of a patient feedback document is also fixed: it is written from the author's viewpoint. News stories may feature interviews with many people, who may give differing versions of a story, hence perspective analysis is required in such cases, or question answering technology to answer the question: who said what? Similarly, a movie review may contain discussions or snippets from a film that convolute the overall review (Scheible & Schütze, 2013). Such tasks are beyond the scope of this thesis. We instead focus on the analysis of a single reviewer's viewpoint in the clinical domain.

3.1.2 Sources for online patient feedback

In this thesis, we build upon the definition of the Picker Institute (2015) and define an item of patient feedback as a text-based document that contains the opinions, judgements and suggestions of a patient or family member regarding their experiences with a given health provider. We also restrict the location of the patient feedback to be online, although it is clear that paper-based alternatives do exist. As the topic of these review sites is quite specific to patient feedback, their number is somewhat limited but large enough for distinctions between the sites to be made. We briefly describe three prominent sources for patient feedback:

- **Health-provider review site:** Health-providers such as the NHS have set up websites (http://www.nhs.uk) that enable users to look up their symptoms and get general medical advice. A layer of interactivity has been added to these sites that allow users to post comments describing their experiences and rate their time with the NHS. These are aggregated by practice, and lead to a summary of the service being given. As the NHS is publicly funded, the data is available for download under an Open Government License[1] that enables the data to be adapted and used for research purposes.

- **Patient review sites** A patient review site is a website whose main purpose is to act as a platform to collect and publish patient feedback with the implication that such a venture is for transparency purposes. Examples of patient review sites are http://www.patientopinion.org and http://www.iWantGreatCare.org and http://www.gp-patient.co.uk. The focus of a patient review site is the submission and viewing of user-generated content which tends to highlight individual stories and patient experiences. Patient review sites also allow interaction with the health providers by sending the stories to the relevant staff, which in turn may generate a response, and furthermore change the patient experience into a more interactive one with a follow-up discourse with a professional.

- **Community support sites:** Online support forums provide a basis for communication that focus on a particular disease (http://csn.cancer.org) or aspect of healthcare, such as

[1] `http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/`

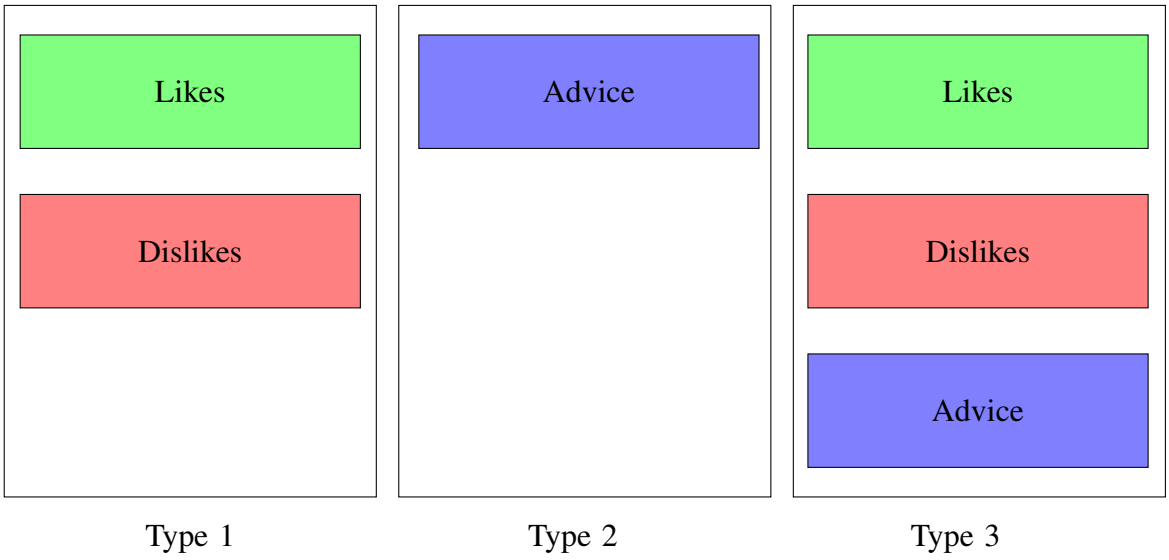


Figure 3.1: Types of patient feedback structure

pregnancy (http://www.mumsnet.com). The focus of these sites is discussion, as opposed to directly reviewing a particular aspect of healthcare. This makes the process of commenting a little less structured than a review submitted through a web form, that limits the accurate mining of opinion posts.

3.1.3 Structure of patient feedback

The format of online patient feedback tends to follow the general structure of online reviews. The general structure is dictated by the fields of an online web form through which a review is submitted. However, as a given, forms will be tailored to the particular area of the health service that the patient is reviewing, so for a hospital patient, a location or specialist may be the topic of the web form. When studying the structure of the files of the NCSD, we found that fields for patient feedback consist of a combination of the following fields: likes, dislikes and advice. Figure 3.1 gives a diagrammatic overview of the structure of patient feedback.

- **Type One** is a two field review, whereby a reviewer lists what they liked in one field, and what they disliked in another. This distinction is useful when distinguishing between aspects of care that have an associated positive or negative sentiment. The structure of comments in these fields may vary. For example, sometimes only a list of aspects is given,

whereas sometimes a full, structured review is given in these fields.

- **Type Two** is a single field review that gives a general overview of the patient's experience. The sentiment of the review should be explicit from what is written in the text, but sometimes a star rating is also given alongside the review. This is not always the case, however, due to multiple sentiments being expressed in this type of review. This type tends to be associated with any advice that a patient may want to give.

- **Type Three** is a three field review; a combination of formats one and two. The likes and dislikes are relative to type one, and the advice relating to type two. Despite being entitled advice, the type two review can be an in-depth analysis of a particular aspect of care or can be used to compare and contrast what was initially detailed in the likes and dislikes, and generally conclude the review.

Each type introduces a different layering of information to the review process which in turn can be used in different aspects of the sentiment analysis process. Type one gives a high-level summary of a patient's opinions, clearly discriminated by the two fields that have distinct, opposing polarities. Terms used in type one reviews which may not have a prior polarity associated with them now convey an implicit sentiment due to the field membership. In comparison, type two reviews typically give a more detailed review than type one, however, the overall document polarity may not always be distinct from the content, and so a star rating may be relied on to infer the overall attitude of a document. This can be problematic, with a 3-star rating either implying an average review, or a review that has examples of polar extreme instances within the document, but averaging does not yield a polar extreme score. Type three provides a holistic approach to sentiment analysis systems. Both types one and two are combined in the type three reviews, and in turn, provides a deeper level for analysis.

3.1.4 Organisation response to patient feedback

Typically the process of leaving a review on a website is an isolated process. A reviewer leaves their feedback about a product on a website such as amazon.com or their review of a film (rot-

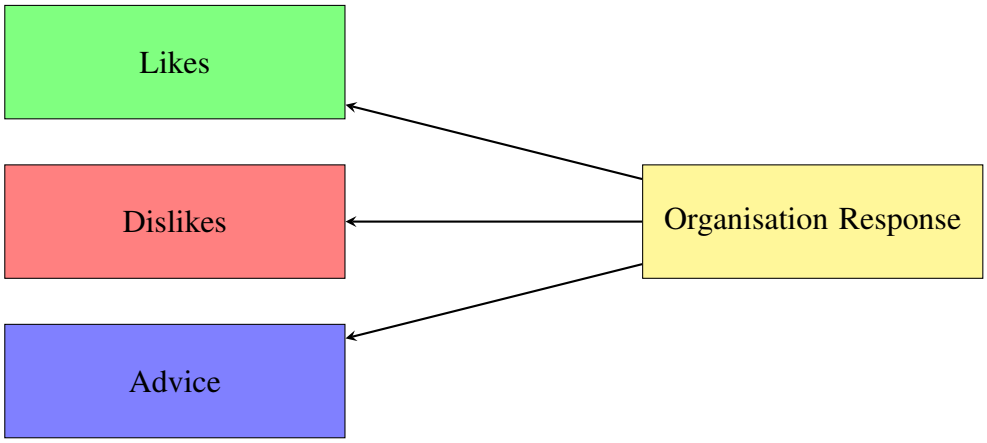


Figure 3.2: The organisational response function in patient feedback

tentomatoes.com) and the site aggregates the scores of the product or film to give it an overall rating. This is fine where the entity is static; that is the product is a physical consumable. However, for a service, a review left in isolation is not adequate. Take for example tripadvisor.com. Hotels can be reviewed on this site, but this is only one side of the multi-faceted review process. Trip Advisor also acts as a portal for hoteliers to respond to the reviews, enabling a feedback loop whereby a hotel indicates that they take customer service seriously (TripAdvisor, 2014).

The response mechanism also forms a significant part of the feedback process on the NHS Choices feedback portal. Figure 3.2 demonstrates the function of the response. It gives the opportunity to respond to a number of aspects of the initial item of patient feedback, whether they be the positive comments, the negative comments, or any advice that the review may be attempting to give. The following response is given in the NCSD:

NHS: *Thank you for such positive feedback. We aim to provide a holistic approach to patient care and are pleased that you have had such a good experience with us.*

Without seeing the feedback that the above response is replying to, it is clear that the review left by the patient expressed a positive sentiment. Similarly, the below comment is indicative of a negative sentiment being expressed in the patient's feedback:

NHS: *I am sorry to hear you are not satisfied with the level of service you have experienced. We have taken on board your comment and have programmed the check in screen to tell you your expected waiting time at the point you check in.*

While not being sources of opinions themselves, the responses reply in such a way that is indicative of the sentiment of the original comments. This mirroring of sentiment provides an alternative source to determine a patient's viewpoint, but this approach has not been considered before in the sentiment classification literature. The context that the response gives will be discussed further in Chapter 5.

3.2 NHS Choices Dataset

In January 2010, the UK Government launched `https://data.gov.uk` as a portal to access open government data. Amongst the 19,000 datasets covering data from a number of government departments, is the NHS Choices dataset[1]. The dataset contains hospital ratings, hospital comments and related responses, and GP comments and responses which are available as three separate XLS files.

Although the overall hospital ratings hold interesting statistical data regarding the likelihood of recommendation for each hospital in the UK, it is the comments and response data that is relevant to this thesis. Each spreadsheet contains a number of rows, each corresponding to a separate item of patient feedback, and a number of columns, representing the metadata of the feedback. Of the metadata, the likes, dislikes, advice and organisation responses were extracted. After extraction, the corpus contained 125, 671 machine readable documents that were converted to UTF-8 text files and stored verbatim in a PostgreSQL database. There are around 11, 750, 000 words in the dataset and the total number of word types across the different review types was 67, 749. An overview of the data is given in Table 3.1. All data used in this thesis is available for use by other researchers as a download from `http://www.cs.bham.ac.uk/~pxs697/datasets/`.

[1] `https://data.gov.uk/dataset/england-nhs-nhschoices-organisations-hospitals\-patient-comments`

	Files	Number of word tokens	Number of word types
Type 1 reviews	32,963	2,151,165	30,670
Type 2 reviews	46,433	5,748,409	51,189
Organisation responses	46,275	3,856,557	26,315
Total	125, 671	11,750,131	67, 749

Table 3.1: NCSD statistics

3.3 Annotation

Current datasets for evaluating sentiment classifiers, while valuable to the development of the field, are not applicable to our proposed recalibration framework. Most relevant to our work is the forum data set (Murakami & Raymond, 2010). However, this is too general for the purposes we are examining due to deviation in discourse topic. Therefore, a dataset has been developed for sentiment classification with the related documents that are required for the response recalibration framework. Unlike other online reviews used to investigate the potency of sentiment classification algorithms, this dataset does not contain a user ranking or score to accompany their comment. An annotation phase is therefore required in order to use the documents as an evaluation dataset for our algorithms.

3.3.1 Annotation Process and Schema

The type 2 reviews and the organisation responses were annotated in accordance with the following annotation schema. Each text was first read by the author of this thesis and labelled on the basis of the predominant sentiment of the text as follows below. The annotation process was accommodated through the development of a custom annotation interface programmed in Java to retrieve comments from the PostgreSQL database, and to write the associated labelling choice given by the annotator back to the database. Each comment was revealed to the annotator in an iterative manner and no context was given to the review or the response; only the text in focus was presented for annotation, along with a labelling choice. The annotator would select a labelling by entering either 1 (positive), -1 (negative), 0 (neutral), 2 (mixed positive) or

-2 (mixed negative). The criteria for selecting a particular labelling sentiment is detailed in the following subsection.

Type 2 Review Annotation Scheme

- Positive: Use this annotation if the text seems to be communicating a positive sentiment in the body of the text.
- Negative: Use this annotation if the text seems to be communicating a negative sentiment in the body of the text.
- Neutral: Use this annotation if the text does not seem to be communicating an opinion.
- Mixed Positive: Use this annotation if the text seems to be communicating a mixture of sentiments, however, the predominant sentiment is positive. This document may be comparing and contrasting a number of features in the text, but the conclusion will have a positive stance.
- Mixed Negative: Use this annotation if the text seems to be communicating a mixture of sentiments, however, the predominant sentiment is negative. This document may be comparing and contrasting a number of features in the text, but the conclusion will have a negative stance.

Organisation Response Annotation Scheme

- Positive: Use this annotation if the text appears to be responding to a positive comment that has been submitted to the NHS Choices website.
- Negative: Use this annotation if the text appears to be responding to a negative comment that has been submitted to the NHS Choices website.
- Neutral: Use this annotation if the text appears to either be irrelevant or not be responding to any particular comment that has been submitted to the NHS Choices website.

- Mixed Positive: Use this annotation if the text appears to be responding to a mixed positive comment posted to the NHS Choices website. There may be a mixture of positive and negative entities responded to, but overall the stance of the response will be positive.
- Mixed Negative: Use this annotation if the text appears to be responding to a mixed negative comment posted to the NHS Choices website. There may be a mixture of positive and negatives entities responded to, but overall the stance of the response will be negative.

3.3.2 Annotation results

A subset of 4,059 type 2 reviews and their related responses were annotated with their expressed sentiments, and sentiments that they were responding to, at the document-level. The reviews contained 254,611 words, of which 10,325 were unique. Corresponding responses contained 403,315 words, of which 9,115 were unique. Despite a larger average document size, the response vocabulary was smaller than the comment vocabulary, indicating a possible constraint on the vocabulary used for responses.

An initial pass of the type 2 reviews highlighted that document-level sentiment was not merely a binary positive or negative sentiment, but often weighed up mixed sentiments before giving a conclusion. Due to this observation, the data was initially annotated with the five-class annotation scheme, defined above. This includes neutral, mixed-positive (labelled in the table as *+2*) and mixed-negative categories (labelled in the table as *-2*). The mixed categories denote that varying sentiments are present in the document, but one sentiment is more salient than the other.

Results of this annotation are presented in table 3.4. Given the annotations, the agreement between the categories of the reviews and their responses is calculated using Cohen's kappa coefficient (Cohen, 1960). Between all categories of the reviews and responses $\kappa = 0.4294$, but observing only the positive and negative labellings of the reviews and responses there is an increase in agreement, whereby $\kappa = 0.761$, a good level of agreement. This agreement is indicative of the level to which the sentiment expressed in a comment is mirrored and acknowledged in a related response. The result that skews the agreement statistic is the positively

labelled Type 2 reviews that are responded to in a negative fashion. This can be attributed to a response replying to a comment posited in the dislike section of the Type 1 review. These are not indicative of incorrect labelling however, but a response that responds to a Type 1 review instead. Due to the low agreement between the five classes of the reviews and the responses, the experiments in Chapters 4 and 6 only make use of the positive and negative annotated type 2 reviews and responses, and do not include the neutral or mixed sentiment comments.

The annotation was carried out by the author of this thesis, but to ensure reliable annotations, an inter-rater study was undertaken by a colleague with a computer science background. A subset of 100 type 2 comments annotated by the author were selected with a distribution of 50 positive comments and 50 negative comments only, as these would be the classes for categorisation used in the classification experiments. Agreement between the annotations was high, with no explicit disagreements in labelling, but uncertainties regarding four comment labelled as positive and seven labelled as negative were raised. A common theme of recommendation was found in all eleven comments. For example, one of the positive comments that the rater was unsure about was the following:

I recommend this practice because of the service I have received and from the dentist I am lucky to have found. I cannot judge any other dentist at this practice which I have not had dealings with.

The rater brought up the point that the uncertainty arose from the fact that the positive sentiment in this short type two review was conveyed in quite a neutral manner. The issue of recommendation being a key indicator of positive feedback was discussed, and it was decided that such comments should remain in the dataset with a positive labelling for recommendation. Similarly, it was decided that comments that suggested not recommending a practice or hospital should also be included as negatively labelled reviews. These uncertainties can be viewed as examples of implicitly conveyed sentiments, which despite having been problematic for sentiment classification in the past (Greene & Resnik, 2009), should not be disregarded as they are valid modes of sentiment conveyance in the patient feedback domain.

Sentiment Label	Review Example
Positive	Although I could not attend I was very impressed with the practice 'open house morning' for patients to make suggestions. And even more amazing the fact that so many of the things suggested have been speedily put in place! well done!
Negative	The doctor I saw yesterday was unhelpful and at the time, felt very cruel with a complete lack of empathy and as I've since confirmed, inaccurate in their statements. There were no positive suggestions or help offered to me and I was basically told to go away and live with it. It's not the first time the doctors have been dismissive and unhelpful and I should have known better really and tried to ask for the visiting doctor. I would never recommend this surgery and we shall be moving to another surgery as soon as possible, something we have been considering for some time. I just wish we'd done it sooner.
Neutral	These comments do not relate to a particular visit to the surgery but represent an overall view (which is shared by my husband) as to how this surgery operates.
Mixed Positive	I am very happy with my dentist they are very polite and the care and treatment is good, just a shame about the reception staff.
Mixed Negative	The doctors is an average old fashioned practice and you get adequate health care but beware if you make any kind of complaint the care you get reduces to 'just enough' patient care

Table 3.2: Verbatim sample type 2 reviews with their associated sentiment label.

Sentiment Label	Response Example
Positive	What great comments. We will strive to maintain high standards!
Negative	Thank you for bringing your concerns to the attention of the practice. The practice is committed to providing a high quality, patient-focused service. Complaints and comments from patients are taken very seriously, as we want every patient to feel satisfied with the services we provide. If you would like to contact the surgery we would be happy to look at your concerns in more detail.
Neutral	Thank you very much for your comments which are being carefully considered by the clinical teams.
Mixed Positive	Thank you for taking the time to post a comment, it is always nice to receive positive feedback and is much appreciated. I am sorry that you had to make a second visit but I hope everything is sorted out for you now. However, if you have any problems please do not hesitate to contact me.
Mixed Negative	Thank you for your comments, we are pleased that you find the GP and staff helpful. With regards to online appointments, this is a problem which we are aware of and are working with our clinical system supplier to rectify. As it is their website unfortunately we are dependent on them to fix the issue but we are keeping in close touch with them, this appears to be a national problem.

Table 3.3: Verbatim sample organisational responses with their associated sentiment label.

$S_{Comment}$ \ $S_{Response}$	*-2*	*-1*	*0*	*+1*	*+2*
-2	3	139	6	5	12
-1	8	2,022	92	33	117
0	0	153	25	102	44
+1	4	251	83	671	187
+2	1	68	1	15	17

Table 3.4: Comment-response sentiment label confusion matrix. Category key: -2 = mixed-negative, -1 = negative, 0 = neutral, +1 = positive, +2 = mixed-positive.

3.4 Data Analysis

The patient feedback data studied in this thesis are inherently text-based, and hence methods that are able to study the qualities of a text-based dataset are the subject of this section.

Hunston (2011) proposes that corpus linguistics methods are appropriate for examining the evaluative qualities of a text. While this approach may generalise the subtleties of evaluative content, Hunston demonstrates that it is possible to use the methods associated with corpus linguistics to examine evaluative documents. Corpus linguistics concerns the collection and study of language in electronic texts (McEnery & Hardie, 2011). This form of analysis is therefore well suited to the study of sentiment conveyance in text, where certain words may be used in different contexts, thereby altering the sentiment which they convey. We find this approach particularly appropriate for studying language usage in the NCSD, and therefore in this chapter, we shall apply a number of corpus linguistic procedures to analyse the evaluative qualities of the NCSD.

This process of data analysis using techniques from corpus linguistics will be undertaken using AntConc 3.2.4m (Anthony, 2011). This tool enables the examination of concordance, cluster, collocate, word frequency and corpus keywords. The first stage of the corpus analysis will be to use AntConc to compile frequency lists of the respective sections of the corpus. Following a process of part-of-speech tagging, tag frequency will be discussed. Finally, keyword analysis will be carried out, examining the words that distinguish the NCSD from a standard reference corpus of written British English, the British National Corpus (Leech, 1992), hereafter referred to as the BNC. Although AntConc is able to cluster the data, we will not examine the data using this technique as the domain is already restricted, and general themes can be revealed from the other analyses. Details of the relevant stages of the data analysis will be discussed further in the following subsections.

3.4.1 Frequency analysis of patient feedback

As Partington (1998) notes "the main sort of corpus-based research into lexis using corpora investigates the frequency of words." Frequency analysis is a commonly employed method from corpus linguistics that is used to quantitatively describe a document set and is useful in developing meaning representations wherever the frequency of a term may be an insightful and discriminating factor, such as information retrieval, text categorisation or sentiment analysis. Full text is undoubtedly a far richer and more informative representation of meaning, but the analysis of a collection of complete texts lacks a robust method for straightforward comparison without significant loss of structure of a text (Kilgarriff, 1996b). For this reason, frequency analysis will be a starting point for investigating whether different review types exhibit different modes of sentiment conveyance, potentially making one type of review more appropriate for the sentiment classification of patient feedback. Typically, frequency analysis generates a ranked list of word tokens. The word tokens in the frequency list can consist of a single word, a unigram consisting of a mixture of upper and lowercase alphabetic characters, or clusters of words, typically limited to bigrams and trigrams. The frequency analysis presented in the following sections will focus on only examine unigram frequencies. Bigram frequencies were also examined but were not found to yield any more insightful information about the nature of the data than unigram frequencies alone. After an initial pass of the data, reported in section 3.5, stop words were removed and all characters were converted to lower-case in order to normalise the approach to compiling the frequency lists.

Alongside the tokens reported in a frequency analysis, either observed frequencies or normalised frequencies can be reported. Observed frequencies are the total number of occurrences of a word, whereas the normalised frequency is the number of occurrences of a word per a certain amount of words. We choose a factor of a thousand as the normalising factor with which to calculate and report normalised word token frequencies.

3.4.2 Key word in context analysis

A concordance or key word in context (KWIC) analysis enables the observation of the "company a word keeps" (Baker et al., 2006). Given a search term, a concordance analysis returns a list of strings, centred on the search term, with a context window of words either side of the search term. This enables the context to be sorted and allows a user to view and distinguish a word's usage scenarios. However, this method is problematic due to the possibility of an overwhelming result set when searching for frequent terms in a corpus. Sinclair (1999) suggests limiting the results to a size of 30 lines, analysing this, and repeating until no further word usages can be determined.

In this chapter, KWIC analyses will be interspersed with the other analyses of the data. These will be used to illustrate the nature of particular keywords, and highlight their usage in the patient feedback domain.

3.4.3 Part-of-speech tagging

Following a frequency analysis of the corpus, the distribution of the part-of-speech (POS) tags across the different sections of the corpus will be examined. A difference in part of speech usage could indicate a difference in writing style and a potentially different way of conveying sentiment across review types, just as differences in word patterns can be used to differentiate between speech and written text (Biber, 2009), for example.

As the data is submitted verbatim, the corpus initially contains no part of speech tags. TagAnt is used to tag the documents (Anthony, 2015). To tag each document in the corpus, TagAnt uses the TreeTagger algorithm (Schmid, 1994). TreeTagger is a probabilistic POS tagger, based on the implementation of decision trees to tag ambiguous terms. It is competitive with state-of-the-art taggers, achieving approximately 96% accuracy on data from the Penn-Treebank corpus.

3.4.4 Keyness analysis

While frequency analysis has its benefits in determining the usage of popular terms, high frequency words such as closed class words may not necessarily be representative of the subject of the corpus. Instead, an investigation into the words that are most representative of the content of the documents in a corpus, irrespective of high frequency terms that do not contribute to the overall meaning of a document, should be carried out. A word is 'key' if the frequency of its appearance is found to be significantly higher in the main corpus than in a reference corpus. To calculate and rank the keyness of a word a number of potential significance tests that have been proposed. Amongst these, the log-likelihood and χ^2 tests are suggested in the literature (Rayson & Garside, 2000). The latter has been found to be unreliable at high word frequencies (Kilgarriff, 1996a) or frequencies of occurrence lower than five (Dunning, 1993). The log-likelihood measure is therefore used to calculate a word's keyness. The log-likelihood statistic (LL) is first computed by defining a contingency table of word frequencies. The definition for the calculation of LL as presented by Rayson & Garside (2000) is repeated here for reference.

Table 3.5: Contingency table for corpus word frequencies

	Corpus One	Corpus Two	Total
Freq. of word	a	b	a + b
Freq. of other words	c - a	d - b	c + d - a - b
Total	c	d	c + d

The values of *a* and *b* are referred to as the observed value *O*. Given the observed value, an expected value *E* is calculate as follows, where N_i is the total number of words in corpus i:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i} \tag{3.1}$$

By equating the values from Table 3.5 into the above equation, we may calculated E for each corpus as follows:

$$E_1 = \frac{c \times (a+b)}{(c+d)} \tag{3.2}$$

$$E_2 = \frac{d \times (a+b)}{(c+d)} \tag{3.3}$$

Following the calculation of the expected values, the LL is then calculated as follows:

$$-2\ln\lambda = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right) \tag{3.4}$$

Using this equation and substituting the relevant values of O and E into this, the LL of a given word is then calculated as follows:

$$LL = 2((a \times \ln(\frac{a}{E_1})) + (b \times \ln(\frac{b}{E_2}))) \tag{3.5}$$

The log-likelihood statistic is calculated in this way for each word in the NCSD in comparison to a reference corpus. The resulting LL values are ranked, and the higher scores are indicative of words that are used unusually more frequently than the reference corpus.

The keyness analysis is run with both an external reference corpus, the BNC, and also within the NCSD, using contrasting review types and review polarities as a reference corpus. When comparing to the external reference corpus, the BNC, review types and polarities are compared to the written BNC frequency list, compiled by Laurence Anthony. This consists of approximately 334, 660 word types and 85, 887, 272 word tokens, spanning texts from a number of genres that are deemed to be representative of written British English. As our corpus is of a specific genre that is not necessarily represented in the BNC, this analysis should yield representative terms that are not typically used in everyday written British English, such as clinical terminology, and the analysis should also highlight terms that are particularly related to positive and negative patient feedback.

Despite the potential insights that can be revealed by running a keyness analysis, it is not without its shortcomings. The analysis ranks the keywords by their log-likelihood score, and

so the highest score could be taken to be the most representative word of the corpus under examination. However, if we are specifically examining the sentiment of review documents and find the highest ranking keywords for a subcorpus of documents conveying a particular sentiment, it would be unwise to assume that the sentiment is the only one that a given keyword may ever convey. The keyness analysis should be treated as indicating the words that have a higher association with a particular sentiment than the other. However, we cannot dismiss the fact that in certain contexts, the opposing sentiment may still be conveyed.

3.5 Frequency analysis

3.5.1 Results: Type 1 reviews

The most frequent unigrams used in patient feedback are first examined. Frequency analysis is performed on the 32,963 Type 1 reviews, consisting of 2,151,165 tokens of which 30,670 are unique.

Table 3.6 gives an overview of the most frequent words used in all Type 1 reviews. What is immediately obvious is the number of closed class terms that offer little information characterising information about the dataset as a whole. There are four entities mentioned in this table: *staff*, *time*, *appointment* and *hospital*. These could be seen as generalising the comments, but on the surface, these do not convey a sentiment that can be gleaned about the dataset. In fact, the only token that does have sentiment-bearing connotations in the most frequent terms of Type 1 reviews is the token *care*; however, it is unclear whether this is used as a noun or a verb. To gain a better insight, a frequency list of the posts tagged as positive or negative would better describe the sentiment-bearing assets of the Type 1 reviews.

Table 3.7 shows the most frequent tokens from the positive and negative dataset subsets, respectively. In an initial construction of this list, a number of high frequency terms that conveyed little about the sentiment of the reviews, such as *the* and *they*, were present. Due to this, a stop-word list is used to filter the frequency list. Stop words may play vital roles in the structure of a sentence, however, they do not give much as to the way of the topic or sentiment of a document.

Therefore, to clarify what the general topics of the positive and negative topics of the Type 1 review are, all stopwords are removed. The stoplist used is a list of 571 common terms from the SMART information retrieval system (Buckley, 1985). This forms a basis for stopwords used in the Rainbow (McCallum, 1996) system that in turn is the stop list used in Weka, that are used for the machine learning experiments in this thesis. In order to use this stoplist, all data was treated as lower case, and normalised frequencies per thousand tokens are in turn calculated in respect of all tokens being treated as lower case also.

In Table 3.7 it is striking to see the similarities between the most frequent unigrams of the positive and negative Type 1 reviews. From the positive word list only the unigrams *friendly, treatment, good, helpful, treated, excellent, nurses, made, service, professional, feel* and *dentist* did not appear in the negative word list. Likewise, from the negative wordlist only the unigrams *patient, patients, told, waiting, wait, reception, people, appointments, hours, back, phone* and *times* did not appear in the positive word list. There is an overlap for the unigrams *appointment, care, day, doctor, doctors, gp, hospital, nurse, practice, staff, surgery, time* and *ward*. Of course, due to space limitations, this only shows an analysis of the top twenty-five most frequent terms, and all unigrams do feature in the corresponding sentiment's list, albeit with a lower frequency ranking.

Table 3.6: Top 50 tokens from all Type 1 reviews. Frequencies normalised per 1000 tokens (PTW).

Unigram	Freq.	PTW	Unigram	Freq.	PTW
the	91814	42.681	this	13389	6.224
and	77009	35.799	it	13282	6.174
to	75240	34.976	as	12575	5.846
I	66180	30.765	you	11157	5.186
was	47038	21.866	are	11118	5.168
a	43662	20.297	all	10616	4.935
of	30379	14.122	an	10552	4.905
in	28397	13.201	i	9401	4.370
my	25944	12.060	been	9300	4.323
for	23126	10.750	time	9095	4.228
have	21336	9.918	but	8978	4.174
that	19464	9.048	would	8615	4.005
with	18092	8.410	appointment	7744	3.600
is	17066	7.933	care	7387	3.434
staff	16528	7.683	by	7211	3.352
on	16292	7.574	t	7089	3.295
me	16116	7.492	when	7027	3.267
The	15949	7.414	so	6928	3.221
at	15196	7.064	from	6872	3.195
not	15054	6.998	n	6825	3.173
had	14591	6.783	hospital	6758	3.142
very	14326	6.660	could	6503	3.023
be	14099	6.554	there	6423	2.986
they	14049	6.531	who	6245	2.903
were	13835	6.431	about	5710	2.654

Table 3.7: Top 25 unigrams from positive and negative Type 1 reviews. Frequencies normalised per 1000 tokens (PTT). Italicized terms are unique to a particular sentiment in these lists.

POSITIVE			**NEGATIVE**		
Unigram	Freq.	PTT	Unigram	Freq.	PTT
staff	12570	28.666	staff	4974	14.609
care	5762	13.140	appointment	4434	13.023
hospital	5280	12.041	time	4328	12.711
time	4813	10.976	*patients*	3341	9.812
ward	3796	8.657	*told*	3110	9.134
friendly	3777	8.613	doctor	2806	8.241
appointment	3429	7.820	hospital	2660	7.812
surgery	3416	7.790	*waiting*	2627	7.715
doctor	3374	7.694	surgery	2151	6.317
treatment	3242	7.393	day	2046	6.009
good	3191	7.277	*patient*	2015	5.918
helpful	2935	6.693	care	1962	5.762
practice	2795	6.374	ward	1797	5.278
doctors	2789	6.360	nurse	1685	4.949
treated	2630	5.998	*wait*	1630	4.787
excellent	2533	5.777	*reception*	1618	4.752
nurses	2485	5.667	*people*	1615	4.743
day	2403	5.480	gp	1577	4.632
nurse	2324	5.300	*appointments*	1551	4.555
made	2302	5.250	*hours*	1538	4.517
service	2208	5.035	doctors	1534	4.505
professional	2186	4.985	practice	1525	4.479
feel	2141	4.883	*back*	1507	4.426
dentist	1902	4.338	*phone*	1375	4.038
gp	1889	4.308	*times*	1353	3.974

3.5.2 Results: Type 2 reviews

Similar to the Type 1 frequency analysis, the removal of stopwords is more revealing of overall comment sentiment. In Table 3.10 *good, work, excellent, feel, great, recommend, experience, nurses, ward* and *friendly* all feature highly in the positive Type 2 comments, whereas the words *appointment, told, gp, back, day, waiting, pain, make, reception* and *left* are all unique to the twenty-five most frequent unigrams used in the negative Type 2 reviews.

The list of unique, high-frequency terms is somewhat similar to that of Type 1. This could suggest similarities in word usage, which would indicate that training a sentiment classifier on reviews of either type may not be detrimental to model generation due to the similarities in vocabularies. However, there are subtle differences that appear. For example, *recommend* appears frequently in the positive comments of Type 2. These types of comments are free-form by nature and are not restricted to only discussing a patient's positive or negative viewpoints. A recommendation is used to indicate the best course of action. While it appears 6.568 per thousand tokens in the Type 2 dataset, it only appears 0.434 times per thousand tokens in the Type 1 positive dataset, indicating the more descriptive and less suggestive nature of Type 1 reviews.

ell looked after. I coudldn't	recommend	it more. It made the experien
my knowledge to which i would	recommend	someone. Sorry for bad punctu
clean. I would wholeheartedly	recommend	the SCBU facility."
thanks to all and I certainly	recommend	this hospital to all. John B
d with dignity and respect. I	recommend	the Cobalt for any procedure
d my stay of 8 days and would	recommend	this part of the hospital to
theatre staff. I would highly	recommend	having a child here they are
nd comfortable life. I would	recommend	St. Albans hospital to all my
provided. I would definately	recommend	F4 and the preoperative care
g to put me at ease. Cannot	recommend	highly enough."
judgement was sound. I would	recommend	her to anyone. Once our bab

Table 3.8: A sample of concordance lines for *recommend*

Table 3.9: Top 50 tokens from all Type 2 reviews. Frequencies normalised per 1000 tokens (PTT).

Unigram	Freq.	PTT	Unigram	Freq.	PTT
the	239990	41.749	you	30826	5.363
to	207066	36.021	very	30755	5.350
I	196765	34.229	The	29820	5.188
and	196387	34.164	were	29223	5.084
a	122479	21.307	an	29105	5.063
was	120828	21.019	are	26874	4.675
in	76703	13.343	all	26109	4.542
of	75924	13.208	but	25818	4.491
my	73692	12.820	been	25760	4.481
for	64426	11.208	would	23551	4.097
have	58981	10.260	t	22839	3.973
that	54090	9.410	n	21664	3.769
with	47788	8.313	time	21449	3.731
had	44731	7.781	appointment	20558	3.576
is	44195	7.688	so	20443	3.556
me	43946	7.645	i	20376	3.545
on	43562	7.578	by	19796	3.444
at	41356	7.194	from	18661	3.246
not	39240	6.826	care	18534	3.224
they	38497	6.697	hospital	17876	3.110
this	38490	6.696	there	17223	2.996
it	37235	6.477	when	17156	2.984
be	35000	6.089	who	16808	2.924
as	33469	5.822	no	16740	2.912
staff	32314	5.621	do	16136	2.807

Table 3.10: Top 25 unigrams from positive and negative Type 2 reviews. Frequencies normalised per 1000 tokens (PTT).

POSITIVE			**NEGATIVE**		
Unigram	Freq.	PTT	Unigram	Freq.	PTT
staff	386	26.687	*appointment*	838	12.984
hospital	255	17.630	time	750	11.621
care	206	14.242	doctor	745	11.543
practice	185	12.790	*told*	708	10.970
good	181	12.514	surgery	707	10.955
surgery	145	10.025	staff	677	10.490
service	124	8.573	practice	575	8.909
time	111	7.674	patients	550	8.522
work	109	7.536	hospital	522	8.088
excellent	107	7.398	*gp*	455	7.050
feel	105	7.259	dentist	372	5.764
great	105	7.259	care	362	5.609
patients	102	7.052	patient	344	5.330
doctors	97	6.706	doctors	339	5.253
recommend	95	6.568	*back*	333	5.160
experience	93	6.430	*day*	323	5.005
treatment	93	6.430	treatment	318	4.927
doctor	92	6.361	service	312	4.834
nurses	92	6.361	people	306	4.741
people	89	6.153	*waiting*	299	4.633
ward	85	5.877	*pain*	295	4.571
patient	84	5.808	nhs	287	4.447
nhs	83	5.738	*make*	266	4.122
dentist	81	5.600	*reception*	266	4.122
friendly	75	5.185	*left*	262	4.060

3.5.3 Results: Feedback responses

As well as the different types of feedback that are present in the NCSD, the frequency of words in the responses is calculated. By observing the results of a frequency analysis of the responses, the role of the response in the patient feedback process is characterised.

Table 3.11 lists the fifty most frequent words used in the responses. The ranking list consists of a number of frequently used stopwords, but content words such as *staff, patients, feedback* and *care* that are representative of the clinical domain of the reviews also rank highly in the list. Words that differentiate the ranked list to those of the Type 1 and Type 2 reviews are the words *thank, feedback, sorry* and *contact*. These mark the interactive nature of the response, thanking the user for leaving feedback, apologising for a negative experience, and encouraging further contact with the user.

Table 3.12 show the most frequent words from the subset of annotated positive and negative responses. When a positive comment is received, the responses appear to mirror the tone of the comment. It does so by using the terms: *positive, kind, pleased, happy* and *good*. When some terms are isolated from their context, the term appears questionable in the high-frequency list. For example, a KWIC analysis of the term *taking* displays that it is frequently collocated with the phrase *the time* to the right and *thank you for* to the left. If the term *passed* is considered, without context it seems peculiar, but a KWIC analysis shows that this is typically used in the phrase *passed on*, when indicating that an item of feedback has been sent to the appropriate staff member to notify them that their good work has been recognised.

In the negative frequency list, the term *appointment* features regularly. This potentially highlights the relevance of the response in responding to what has initially been complained about. Relevance is an important characteristic of responses, and we shall build upon this insight in developing a method to improve sentiment classification in the presence of a relevant response.

Table 3.11: Top 50 tokens from all responses. Frequencies normalised per 1000 tokens (PTT).

Unigram	Freq.	PTT	Unigram	Freq.	PTT
to	185052	47.984	will	26118	6.772
the	163750	42.460	patients	26085	6.764
you	119949	31.103	feedback	22375	5.802
and	99117	25.701	time	22110	5.733
your	79685	20.662	very	21062	5.461
for	68476	17.756	can	20539	5.326
that	62549	16.219	not	19347	5.017
of	53978	13.996	staff	18569	4.815
we	52670	13.657	as	18177	4.713
a	51680	13.401	us	17926	4.648
are	50449	13.081	experience	16784	4.352
our	48509	12.578	like	16000	4.149
We	45432	11.780	practice	15514	4.023
have	43512	11.283	contact	15468	4.011
on	38547	9.995	patient	14891	3.861
with	38411	9.960	so	14825	3.844
in	37969	9.845	service	14782	3.833
I	36840	9.553	or	14724	3.818
is	36182	9.382	do	14619	3.791
this	33058	8.572	care	14387	3.731
be	30006	7.781	it	14228	3.689
comments	29355	7.612	all	14203	3.683
Thank	29164	7.562	sorry	14089	3.653
would	28104	7.287	about	13943	3.615
at	26119	6.773	been	13236	3.432

Table 3.12: Top 25 unigrams from positive and negative responses to patient feedback. Frequencies normalised per 1000 tokens (PTT).

POSITIVE			**NEGATIVE**		
Unigram	Freq.	PTT	Unigram	Freq.	PTT
comments	674	48.784	patients	2622	21.340
feedback	428	30.978	practice	2575	21.016
positive	392	28.372	patient	2493	20.347
time	371	26.852	comments	1750	14.283
staff	324	23.451	service	1506	12.292
taking	311	22.510	*contact*	1483	12.104
kind	300	21.714	experience	1308	10.676
experience	275	19.904	feedback	1204	9.827
practice	265	19.180	time	1177	9.606
team	250	18.095	nhs	1155	9.427
care	223	16.140	manager	1086	8.864
pleased	211	15.272	staff	1079	8.806
patient	200	14.476	*appointments*	1028	8.390
patients	196	14.187	care	992	8.097
service	184	13.318	discuss	956	7.803
hospital	183	13.246	*surgery*	911	7.435
passed	159	11.508	*appointment*	901	7.354
hear	155	11.219	*concerns*	797	6.505
provide	126	9.120	team	743	6.064
manager	125	9.048	provide	657	5.362
received	125	9.048	hospital	654	5.338
happy	124	8.975	*improve*	624	5.093
nhs	117	8.468	*day*	620	5.060
comment	115	8.324	hear	597	4.873
good	114	8.251	*pals*	583	4.759

3.6 Part-of-speech analysis

From the results of the frequency investigation for the positive and negative Type 1 reviews, it is apparent that different parts of speech have sentiment-bearing connotations. For example from the positive reviews, the adjectives *friendly, good, helpful, excellent* and *professional* are all indicative of a positive sentiment. There are no frequent adjectives in the negative frequency list, but nouns make up the core of the unique unigrams in the twenty-five most frequent list, for example, *patient, reception, hours, phone* and *times*. We, therefore, examine the frequency of part-of-speech tags in the dataset to examine this further.

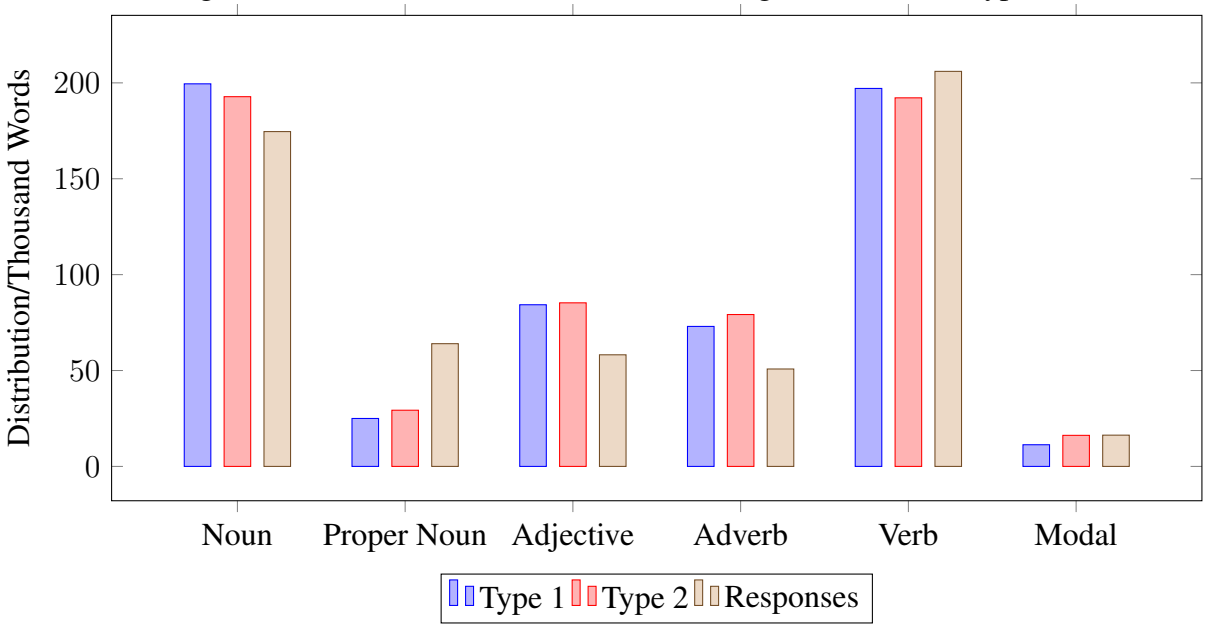
As both Hunston (2011, p. 3) and Liu (2010, p. 26) note, sentiment-bearing words often fall into the categories of adjectives and adverbs. Given this reasoning, the frequency of these parts-of-speech in the respective review types in patient feedback is observed. Nouns may have connotational associations with particular sentiments (Feng et al., 2013), and verbs have also been found to be indicative of emotion in text (Simančík & Lee, 2009), so we also examine these. Modal verbs exhibit qualities of speculation and suggestion, which in turn affect the overall sentiment label of a document, so the distribution of these in the dataset is also observed.

Figure 3.3 displays the distribution of the aforementioned parts-of-speech across the sub-sections of the NCSD. The distribution is normalised per thousand tokens to aid comparison. Comparing only Types One and Two, the distributions seem quite similar. However, Figure 3.4 shows that between sentiments, there are differences between the part of speech distributions. It appears that proper nouns and adjectives are more frequent per thousand tokens in positive reviews, but nouns, adverbs, verbs and modals are more frequent in negative feedback. This will be explored in more depth in the following sections by carrying out a frequency analysis of specific parts of speech interspersed with KWIC analyses.

3.6.1 Adjective distribution

Adjectives are traditionally studied in the literature as the primary source of a document's polarity (Hatzivassiloglou & McKeown, 1997). However, they are notoriously domain and context

Figure 3.3: Normalised distribution of POS tags across review type

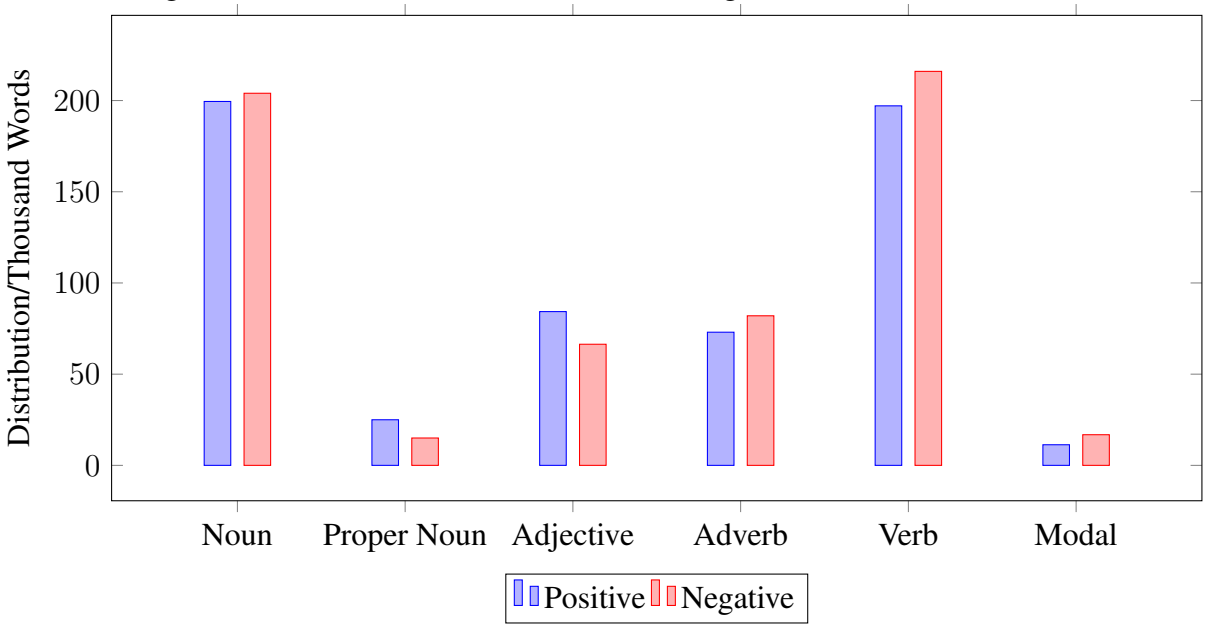


dependent. For this reason, Table 3.14 shows the most frequent positive and negative adjectives across Type 1 and Type 2 reviews. Between the review types, the lists tend to mimic each other, so from observation of a frequency list alone, a conclusion cannot be drawn that there are any adjectives that may be associated strongly with a particular review type.

There are examples in the table of an adjective frequently appearing in both the positive and negative reviews of Type 1. This list includes *good, first, other, same, medical, able, many, last* and *next*. Only good seems to have explicit positive connotations, whereas the other adjectives are seemingly neutral until the context is observed. However, its use under both contexts is interesting. In the positive reviews *good* tends to be collocated with *work* and *very*. In negative reviews, good is negated with *not* and collocates with *service* and *practice*. Here the adjective surrounding context of good effects the overall sentiment that is conveyed. The unique words are better indicators of sentiment: for example *friendly, helpful* and *excellent*, and *rude, long* and *difficult*.

The most frequent negative adjectives do not appear to be overwhelmingly negative. *Wrong, rude, difficult* and *busy* stand out as negative terms, whereas the terms *other, more* and *same*, for example, require a context to determine the negative sentiment that they are helping to convey. *Other* is used in the context of negative reviews from *other patients*, and *more* is used when

Figure 3.4: Normalised distribution of POS tags across review sentiment



requesting improvements: i.e *the receptionist could have been more friendly*. *Same* refers to a negative experience re-occurring and the patient being unhappy with this.

ddressed could make it into a	good	one. I hope someone will reco
understanding and best of all	good	humour. I attended A& E
Keep up the	good	work!
ring visiting hours are not a	good	idea. It clearly stated that
sured that they'll be in very	good	hands
be worried. I found it a very	good	experience. Every member of s
Please dont ever lose the	good	focus that you have it makes
ess it makes. This is not a	good	image when entering the build
Not a	good	practice at all. Wasn't appro

Table 3.13: A sample of concordance lines for *good*

Table 3.14: Most frequent positive adjectives across reviews

POSITIVE		**NEGATIVE**	
Type 1	Type 2	Type 1	Type 2
friendly	good	other	other
good	great	more	more
helpful	excellent	rude	rude
excellent	helpful	good	good
professional	friendly	same	same
first	best	first	medical
great	other	next	new
other	happy	long	last
caring	many	available	first
clean	professional	able	many
efficient	first	only	wrong
polite	medical	many	due
best	wonderful	medical	next
happy	more	new	long
same	lovely	better	able
fantastic	able	difficult	bad
medical	grateful	due	old
able	big	busy	few
much	same	last	different
many	fantastic	different	better
last	much	few	available
lovely	possible	wrong	private
pleasant	last	poor	poor
nice	amazing	least	own
next	brilliant	own	patient
whole	caring	patient	several

Thanks to all the	staff	from the receptionists, nursi
om the receptionists, nursing	staff	, surgeon and anthestist as we
as well as all the background	staff	for making it work so well.
is a big thank you to all the	staff	involved in their care : Card
was very efficient. All the	staff	I encountered were friendly a
ents being slightly deaf, the	staff	seem to think by shouting at
solute madness of making what	staff	there is there redundant, the
ing with a sanity bullet. The	staff	do nothing but moan and moan
de and egotistical members of	staff	just to top it off. its not a
early two months Very rude	staff	who serve food

Table 3.15: A sample of concordance lines for *staff*

3.6.2 Noun distribution

The most frequent nouns are reported in Table 3.16. The table shows that irrespective of review type or review sentiment there is an overlap in the most frequent nouns that are used. For example, Table 3.15 shows concordance lines for the entity *staff*. This noun displays this property; being the most frequent noun for both positive review types, and for Type 1 of the negative reviews, and it ranks fifth in the Type 2 negative reviews.The table shows that it is the surrounding context of the term *staff* that is more indicative of sentiment. This example can also be expanded to cover other nouns in the data. Given this common usage of nouns irrespective of review type or review sentiment, this confirms the assumption that it may not be advisable to observe noun usage to automatically determine the sentiment of a document.

Table 3.16: Most frequent nouns across reviews

POSITIVE		**NEGATIVE**	
Type 1	Type 2	Type 1	Type 2
staff	staff	staff	appointment
care	hospital	appointment	time
time	care	time	surgery
hospital	practice	patients	doctor
appointment	surgery	doctor	staff
surgery	service	hospital	patients
treatment	time	surgery	practice
doctor	patients	day	hospital
ward	treatment	Nothing	GP
practice	experience	care	dentist
doctors	work	people	treatment
day	people	nurse	day
nurses	doctors	ward	service
service	nurses	hours	people
nurse	dentist	patient	doctors
experience	NHS	appointments	pain
dentist	doctor	GP	care
GP	GP	practice	NHS
team	day	reception	appointments
patients	team	times	nurse
times	years	treatment	patient
reception	patient	pain	reception
years	ward	room	i
amp	thanks	doctors	hours
way	appointment	phone	phone
patient	anyone	nothing	way

3.6.3 Verb distribution

The twenty-five most frequent verbs are reported in Table 3.18. Given the descriptive nature of a review, *was* is the most frequent verb across all review types and sentiments. There is little variation between the lists to differentiate review type and associated sentiment.

Among the most frequent verbs of the positive Type 2 reviews are *recommend* and *impressed*. These may be indicative of the free-form nature of Type 2, enabling users to discuss in depth their opinions and advice. Positive Type 2 reviews also feature both *Thank* and *thank* as high-frequency verbs. *Thank* does not appear in the positive Type 1 reviews, however. Interestingly, while positive Type 1 reviews essentially denote what a patient liked or disliked, this also appears to extend to Type 2 reviews, with the inclusion of the verb *like*. In contrast with this, dislike does not appear in the negative Type 2 reviews, indicating that perhaps more subtle mechanisms are being utilised when describing the actions associated with negative feedback. In fact, it is difficult to deduce a negative connotation from observing the negative verbs alone. *Waiting* is perhaps the closest to a clear negative sentiment, with the implication that somebody was waiting for something to occur.

The verb that appears to be the key to revealing a patient's sentiment is *felt*. In the feedback, the pattern `felt JJ` where JJ belongs to a list of polarity bearing adjectives. Concordance examples are given in Table 3.17.

me know what was going on. I	felt	so relieved when it was done
e, they came to visit and you	felt	cared for, and that you could
se of the care I received. I	felt	very well looked after and kn
Consultants were brilliant i	felt	really well looked after. Eve
I moved hospitals because I	felt	with such a strong team of ex
interest in me as a person. I	felt	like an inconvenience. Actual
tually, it wasn't just that I	felt	like a burden, when I suggest
it was my fault alone that I	felt	so uncomfortable. I should
ormed and making decisions. I	felt	I was left uninformed for a g
nship between ward staff . It	felt	more that people were unwilli

Table 3.17: Concordance lines for *felt*

Table 3.18: Most frequent verbs across review types

POSITIVE		**NEGATIVE**	
Type 1	Type 2	Type 1	Type 2
was	was	was	was
were	have	is	is
have	is	be	be
is	are	had	have
had	be	are	had
are	were	have	are
be	had	have	been
been	been	been	have
have	have	were	were
am	am	get	get
has	Thank	did	am
did	has	waiting	do
treated	do	do	see
get	recommend	see	has
see	feel	told	told
seen	thank	being	did
being	done	has	go
do	being	am	being
made	did	go	said
given	like	given	waiting
feel	get	seen	told
waiting	see	said	do
went	seen	do	going
go	treated	wait	make
say	go	told	went
felt	impressed	make	s

3.7 Keyness Analysis

In this section, the results of a keyness analysis carried out on the NCSD are given. A word is key if it is found to occur significantly more frequently in the main dataset in comparison to its frequency value in a reference dataset when using the log-likelihood measure. There is a significant difference in the frequency of words in two corpora to a significance of $p < 0.0001$ where a log-likelihood value > 15.13. In the tables of this section, only the twenty-five words with the highest log-likelihood values are displayed that indicate what the most discriminating terms could be.

Table 3.19 summarises the keyness analyses that are carried out with respect to the main and reference corpora. First, keywords are determined in comparison to the BNC reference corpus. The BNC is a hundred-million word corpus of both written and spoken contemporary British English. Using this corpus should distinguish the words that are used unusually frequently in the patient reviews in comparison to a corpus of general English usage. Next, the keywords found between different review types of the same sentiment are detailed in order to determine any significant differences between reviews and the way that sentiment is conveyed using the different types. Finally, a keyness analysis is carried out using a dataset of one sentiment as the main dataset, and the opposing sentiment's dataset as the reference; for example, the main dataset will be the positive reviews, and the reference dataset the negative reviews. This is used to determine any words that may be significant indicators of sentiment. In doing so, this should improve upon the results of the frequency analysis alone, as these were not satisfactory in yielding discriminating terms.

3.7.1 Comparison with the BNC

Results of a keyness analysis, using the BNC as the reference corpus are shown in Tables 3.20 and 3.21. Results indicate that the topic of the NCSD is indeed healthcare, indicated by key terms *NHS*, *doctor* and *treatment*, amongst others.

The domain is further reinforced by overlapping key terms between the positive and nega-

Table 3.19: Main and reference corpora used in the keyness analyses

Main Corpus	Reference Corpus
Positive T1 & T2	BNC
Negative T1 & T2	BNC
Type 1	Type 2
Type 2	Type 1
Positive T1 & T2	Negative T1 & T2
Negative T1 & T2	Positive T1 & T2

tive review types: these include *hospital, ward, surgery, GP* and *appointment*. The overlapping words *I, my, me* and *was* are also found to be key in comparison to terms of the BNC, which indicates that irrespective of sentiment, patient reviews are written from a first-person perspective, and describe a patient's personal experiences.

The remainder of the keywords that are discovered when comparing to the BNC that do not overlap are good indicators of review sentiment in the clinical domain. For example, the positive keywords *friendly, care, helpful, treatment, excellent* and *thank* are all good indicators of a positive document sentiment. At first sight, there are some words in the negative unique keyword list that may not make sense without observing the keyword in context. For example the word *reception* may appear to convey no sentiment out of context; however, it is frequently collocated with *staff* and *area*, followed by a negative description of it. Also, when the word *told* is used, it tends to describe a situation where a patient was told something incorrect, that caused them great confusion or annoyance. These words are not typical to a sentiment lexicon, however, nor a more subtle connotation lexicon.

3.7.2 Type 1 versus Type 2 keyword analysis

A keyness analysis between Type 1 and Type 2 reviews (Table 3.22) exposes the tendency for Type 1 reviews to focus on objects and their descriptions, such as *clean ward* and Type 2 reviews to discuss the people involved in the patient feedback process: *I, him, her*.

Table 3.20: Over-represented words in all positive comments, calculated using the log-likelihood ratio with respect to the BNC reference corpus.

Type 1			**Type 2**		
Keyword	Freq.	Log-likelihood	Keyword	Freq.	Log-likelihood
i	44550	79727.743	i	1507	2773.538
staff	12570	64766.269	staff	386	2099.707
my	18595	53297.619	my	606	1766.12
very	11512	29009.521	thank	209	1419.056
hospital	5280	23391.265	hospital	255	1383.33
ward	3796	23040.377	surgery	145	1108.121
friendly	3777	21707.133	very	353	860.465
surgery	3416	21437.245	practice	185	825.225
care	5762	21330.321	care	206	825.017
was	29066	21043.436	dentist	81	801.828
me	9882	20719.246	recommend	95	733.235
appointment	3429	19343.712	nurses	92	633.681
helpful	2935	17452.203	gp	74	596.414
doctors	2789	14843.329	doctors	97	572.304
nurses	2485	14787.977	excellent	107	571.592
doctor	3374	14699.808	nhs	83	544.716
dentist	1902	14173.219	ward	85	517.516
nurse	2324	12690.731	helpful	73	446.016
gp	1889	12512.402	n	98	443.82
treatment	3242	12506.677	thanks	81	430.041
treated	2630	11685.894	friendly	75	404.244
excellent	2533	11592.581	me	244	400.19
thank	2435	11293.413	doctor	92	390.261
n	2689	11101.667	all	366	371.066
caring	1880	10891.574	have	481	364.61

Table 3.21: Over-represented words in all negative comments, calculated using the log-likelihood ratio with respect to the BNC reference corpus.

Type 1			**Type 2**		
Keyword	Freq.	Log-likelihood	Keyword	Freq.	Log-likelihood
i	31031	48637.137	i	7232	13564.857
appointment	4434	28418.644	my	2467	6543.591
my	10854	25398.228	appointment	837	5829.309
n	4253	22444.969	surgery	708	5326.638
staff	4974	20196.297	n	853	4818.988
surgery	2151	13081.18	doctor	746	3906.263
doctor	2806	12513.993	gp	457	3787.902
patients	3341	12327.193	dentist	372	3474.001
waiting	2627	11606.488	me	1446	2923.269
gp	1577	10779.656	staff	676	2397.194
me	6258	10750.433	practice	574	2101.259
appointments	1551	10017.56	told	708	2032.907
hospital	2660	9806.674	am	654	2010.46
ward	1797	9735.283	hospital	522	1993.812
reception	1618	9383.158	patients	551	1929.882
nurse	1685	9085.236	doctors	339	1788.303
patient	2015	8195.905	appointments	259	1764.226
quot	889	7967.782	receptionist	201	1721.355
was	18209	7950.98	nhs	285	1675.507
told	3110	7911.559	rude	215	1618.345
nothing	2929	7662.292	they	1984	1571.554
dentist	1014	7428.439	reception	265	1565.917
receptionist	998	7395.706	quot	123	1478.412
rude	1093	7389.251	have	2101	1408.868
doctors	1534	7362.971	patient	348	1390.715

There is a high proportion of adjectives in Type 1 keyword list. *Friendly, helpful, professional, clean, exceptional, efficient, caring, polite, nice, pleasant, modern, good* and so on are all highly ranked when comparing the Type 1 dataset to Type 2. There is also a high proportion of entities mentioned in Type 1 reviews, which is to be expected as the aspects that patient's liked and disliked are distinctly requested in this type of review field. For example, *staff, reception, attitude, practice, parking, communication, midwives*, and *facilities*.

A number of key pronouns are found in Type 2 reviews. *I, she, he, her, we, his, him,* and *my* all demonstrate the more personal approach of Type 2 reviews, as opposed to the aspect based content present in Type 1 reviews. The descriptive element of Type 2 reviews is shown through the high use of terms relating to time and recency of an ailment or treatment: *recently, weeks, later, today, last, years, ago*.

The comparison of the top twenty-five keywords for Type 1 and Type 2 reviews shows that Type 1 reviews may communicate sentiment in a more traditional, explicit manner, whereas the Type 2 reviews may communicate sentiment in a more implied manner. A keyness analysis comparing document of positive and negative sentiment should show how this differs between review polarities.

3.7.3 Positive vs negative keyness analysis

Results of the positive and negative review keyness analysis (Table 3.23) bring together what has been discussed so far in characterising the positive and negative instances of patient feedback. The *caring*, *kind* and *professional* nature of the NHS is highlighted in the positive reviews, whereas the *rude* staff and *waiting* times for *appointments* seem to all be aspects of the NHS that *could* or *should* be *improved*, highlighted by the key terms of the negative reviews.

Again, what is highlighted by the polarity-based keyness analysis is the openness of a reviewer to praise a service, but to hedge their criticisms, or frame them in a non-explicit manner. This could be a detrimental factor when using lexicon-based approaches to the sentiment classification of patient feedback, as an in-domain lexicon may not account for the non-explicit way in which a negative sentiment is conveyed in the patient feedback domain.

Table 3.22: Over-represented words in both Type 1 and Type 2 comments, calculated using the log-likelihood ratio with respect to comments of the opposite type.

Type 1			**Type 2**		
Keyword	Freq.	Log-likelihood	Keyword	Freq.	Log-likelihood
nothing	4962	2815.585	she	18774	493.342
staff	17544	1194.276	he	15876	414.171
friendly	4006	923.356	i	217141	297.523
improved	782	519.976	said	8157	265.27
very	15102	507.298	her	13961	260
were	13842	502.028	had	45450	229.516
liked	434	484.593	told	15894	221.452
helpful	3350	403.666	we	20816	214.791
everything	1991	402.123	now	8418	204.453
clean	1657	378.75	t	22972	197.121
the	107784	324.767	review	1185	169.104
reception	3330	303.599	recently	2908	165.542
patients	5003	272.728	reviews	816	164.517
treated	3214	246.744	will	8879	159.83
professional	2406	223.199	then	12523	151.413
polite	1203	221.745	n	21917	150.798
food	1392	218.742	again	7926	147.363
attitude	1067	212.148	back	9760	146.72
exceptionally	299	200.814	him	5213	145.87
waiting	4023	195.064	his	7409	136.59
environment	424	194.084	another	6943	131.639
efficient	1241	179.717	weeks	6555	128.637
exceptional	491	177.859	has	13332	125.519
times	2890	168.037	my	84662	118.097
communication	798	162.443	went	7556	116.784

Table 3.23: Over-represented words from both positive and negative comments, calculated using the log-likelihood ratio with respect to comments from the opposing sentiment forming the reference corpus.

Positive			**Negative**		
Keyword	Freq.	Log-likelihood	Keyword	Freq.	Log-likelihood
and	52699	4022.591	not	11358	1652.12
friendly	3852	3489.464	be	10096	1469.781
very	11865	3280.243	told	3818	1118.49
thank	2644	2656.372	it	10951	998.844
all	9493	2521.107	patients	3892	972.095
staff	12956	2471.725	t	5236	861.398
excellent	2640	2104.785	n	5106	858.498
helpful	3008	1890.355	should	2444	763.309
professional	2240	1646.62	no	4524	760.893
were	9954	1597.144	or	3799	744.674
caring	1920	1420.515	that	12674	686.424
care	5968	1418.928	waiting	2926	630.364
well	3148	1216.294	more	3061	562.47
always	3129	1200.552	is	10967	551.518
team	1794	1100.142	could	4483	546.869
thanks	1123	1072.161	to	44075	531.942
treated	2678	1004.162	do	4262	526.419
efficient	1171	982.927	if	4106	513.232
ease	1007	958.31	appointment	5271	500.686
kind	1367	911.361	improved	681	453.503
fantastic	1034	886.402	get	4016	447.872
good	3372	864.524	need	2022	423.827
was	29683	855.468	said	1649	404.225
impressed	867	790.725	there	5081	395.988
clean	1425	760.43	rude	1308	390.373

are passed on to the relevant	staff	in general surgery at Stamfor
haring your comments with all	staff	at the hospital. Thank you
rd work and dedication of our	staff	. Your kind comments have been
r feedback to the appropriate	manager	for their attention. However,
ave raised this with the ward	manager	and sister so that your conce
forwarded to the appropriate	manager	for their attention. If y

Table 3.24: Sample concordance lines for *staff* and *manager*

3.7.4 Results of keyword analysis: Feedback responses

Having investigated the language used by different review types and different review polarities, the key terms in the organisation responses are observed. Keywords emerging from a keyness analysis in comparison to the BNC are shown in Table 3.25. Keywords characterising responses that acknowledge comments of both sentiments include the terms: *thank, comments, your, feedback, we, practice, experience* and *patient*. These appear to be used as a matter of protocol, irrespective of sentiment and enable a polite response to be constructed regarding the topic of the feedback.

Positive key verbs include *taking*, *passed* and *hear*, which are often found in the phrases *thank you for taking the time*, *your kind words have been passed on*, and *it's great to hear from you.* The tone of the negative verbs is somewhat more formal, however, advising further contact in order to *discuss* a patient's *concerns* regarding the *service*.

The use of *pleased* in the positive responses indicate the satisfaction in a positive patient experience that is acknowledged in a response, whereas *sorry* indicates a general empathy with a negative scenario that a patient has discussed in their review. A KWIC analysis of staff indicates that positive comments tend to be passed on to the *staff*, whereas negative comments are passed on to a *manager*.

Table 3.25: Over-represented words in both positive and negative comments, calculated using the log-likelihood ratio with respect to the BNC reference corpus.

Positive			**Negative**		
Keyword	Freq.	Log-likelihood	Keyword	Freq.	Log-likelihood
thank	795	7605.362	we	8446	22997.042
comments	674	6764.229	patient	2530	16713.187
your	1336	6559.339	you	8483	15721.483
feedback	428	4892.067	your	4874	15646.656
you	1609	4431.672	our	4217	14575.326
positive	392	2976.903	patients	2642	14253.145
we	920	2461.077	practice	2572	13960.925
our	495	1762.218	comments	1750	12885.204
staff	324	1723.749	feedback	1201	10387.576
taking	311	1658.778	email	857	9256.557
pleased	211	1595.108	sorry	1458	8919.426
kind	300	1553.469	thank	1353	8412.461
practice	265	1422.585	appointments	1026	8280.878
experience	275	1390.74	contact	1480	7897.833
team	250	1271.39	surgery	913	6325.741
patient	203	1253.179	please	1210	6041.022
care	223	974.322	discuss	956	5525.993
patients	198	934.566	appointment	903	5485.031
hospital	184	922.529	service	1509	5179.989
very	351	921.707	experience	1306	4998.691
for	1003	869.998	concerns	796	4830.964
passed	159	858.125	manager	1086	4649.683
hear	155	805.759	to	15418	4395.808
time	371	741.586	website	386	4327.41
regarding	101	736.293	gp	529	3963.248

3.8 Discussion

Prior to the corpus analysis of the NCSD presented in this chapter, the general traits of online reviews in this domain were discussed in section 3.1. In examining the most frequent terms, the distribution of the parts of speech and the key terms of the subsets of the NCSD, the corpus analysis discussed and examined the content, writing style and polarity evoking entities that are characteristic of patient feedback. In doing so, we demonstrate how this domain differs slightly to the traditional sentiment analysis domains such as product reviews.

When considering the writing style of patient feedback, a keyness analysis confirms the assumption that both types of review are written in the first-person. This is demonstrated by the highest ranked keyword, *I*, that is found when calculating keyness in reference to the BNC written corpus. Both *my* and *me* are also found to be key, further reinforcing that the reviews focus on the first-person experience. This is not surprising given the general assumption that a review document aims to give the reviewer's personal opinion, but when comparing the NCSD to Pang and Lee's movie review dataset, despite the movie reviews being from the perspective of the reviewer, *I* is found to occur significantly more frequently in the NCSD than in the movie review dataset.

Common themes in the NCSD are found when comparing the respective review types to the BNC. The entities that are repeated across both the positive and negative review can be grouped into the overarching categories of people, place and action. The people group includes the terms: *staff*, *nurse*, *doctor*, *dentist*, *gp* and *receptionist*. The location group includes: *hospital*, *practices*, *surgery*, *ward*, *appointment* and *reception*. As well as the key entities, we can also group the key terms by the action performance they suggest, and the description of the experience. Actions include: *care*, *service*, *treat*, *told* and *thank*. Positive adjectives include *excellent*, *helpful* and *friendly*, whereas one of the higher ranked negative key adjectives is *rude*.

A comparison of the positive and negative reviews highlights which entities may be discussed more with regard to a particular sentiment. Observing the results of the frequency analysis of nouns in Table 3.16, a high overlap in the most frequent terms is apparent, suggesting that differing opinions about entities are discussed in the feedback. A keyness analysis com-

paring the positive and negative datasets is found to be more discriminative than the frequency list in determining key entities to a given polarity of review. For example, Table 3.23 shows that from the people group, *staff* and *team* tend to be associated with positive reviews; whereas *patients* tend to be associated with the negative reviews. There are no locations among the high ranking key positive terms, but for the negative, the *appointment* appears to be associated with a negative sentiment.

The positive and negative frequency and keyness comparison also highlights the concise, descriptive nature of the positive reviews, in comparison to what appears to be a lack of obviously negative terms in the negative key list. Adjectives dominate the positive keywords: *friendly*, *excellent*, *helpful*, *professional* and so on can all be used to distinguish the positive reviews from the negative. However, verbs such as *told*, *do*, *need*, *get* and *improved*, and the modals *could* and *should*, all rank highly and seem to characterise the negative feedback. This suggests that the way patients structure a review is potentially different to other domains. Patients are more than happy to compliment and do so in an open form. However, negative feedback is given in a more subtle manner, with not so many openly negative terms, but instead, describes actions and suggestions related to the actions.

So, what are the implications for this dataset if it were to be used for training and testing sentiment classification models? First, it is clear that the data is sentiment-bearing, and is therefore suitable for the training and testing of a sentiment classification system. Second, although suitable, sentiment is conveyed in unconventional ways, particularly in Type 2 negative reviews. Given this, a traditional sentiment lexicon such as SentiWordNet (Esuli & Sebastiani, 2006) may not be suitable as a resource for the classification of patient feedback by sentiment for the above traits. A traditional sentiment lexicon contains a list of terms with a score or labelling representing a polarity, so given a review, a lookup process matches words from the document to those in the lexicon, and assigns a score or labelling to the words. A function to map the scores or labels to an overall document sentiment is then applied. The quality of the lexicon often dictates the accuracy of the approach. A general purpose lexicon may provide general coverage, but may be grossly inaccurate when dealing with obscure or niche domains, whereby

sentiment may be conveyed in a non-traditional manner, such as the patient feedback domain. This leads us to the final point, which is that if we are to robustly undertake sentiment analysis in the clinical domain, instead of a lexicon-based approach, a method that learns from the given data may be preferable. For this reason the following chapter examines the applicability of such machine learning methods to classify patient feedback by the sentiment that it conveys.

3.9 Discourse Function

The split in the review types studied in this thesis poses an interesting linguistic problem regarding the effects of the purpose of a review on the use of supervised machine learning classifiers trained for sentiment analysis. The Type 1 review expresses positive and negative aspects of a patient's healthcare, whereas the Type 2 field was used for a reviewer to give their advice regarding their experience. Both offer a different purpose: the likes and dislikes an opinion, and the advice a recommendation regarding the good and bad aspects of their patient experience. Despite this difference in function, both convey a positive or negative sentiment and therefore can be appropriately classified by a computational process. Reviews tend to offer a combination of the two, for example, plot highlights and recommendations in a film review, or aspects and a conclusion in a product review. However, the split enables another aspect of patient feedback to be considered in the sentiment classification: the discourse function.

The purpose of an utterance has long been discussed in the linguistics literature (Wittgenstein, 1953; Austin, 1962; Searle, 1976). We base our definition of discourse function on that proposed by Kinneavy (1969), who argues that the aim of discourse is to produce an effect in the average reader or listener for whom the communication is intended. This could be to share how one is feeling, or perhaps to persuade them. These two discourse functions fall into the *expressive* and *persuasive* categories, respectively. Kinneavy also includes two other discourse functions, informative and literary, in his theory of discourse (Kinneavy, 1971).

To illustrate his theory, Kinneavy represents the components of the communication process as a triangle, with each vertex representing a different role in the theory. This is somewhat sim-

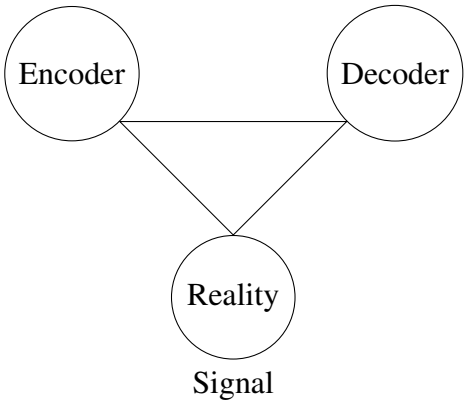


Figure 3.5: The communication triangle (Kinneavy, 1969)

ilar to the schematic diagram of a general communication system that is proposed by Shannon (1948). The three vertices of the triangle are labelled as the encoder, the decoder and the reality of communication. The signal, the linguistic product, is the medium of the communication triangle. The encoder is the writer or speaker of a communication, and the decoder is the reader or listener.

3.9.1 Expressive

In communication, when the language product is dominated by a clear design of the encoder to discharge his or her emotions, or to achieve his or her own individuality then it can be stated that the expressive discourse function is being utilised (Kinneavy, 1971). In this thesis, we take expression to be communicated through text. Since the discourse function is in effect the personal state of the encoder, there is naturally an expressive component in any discourse. We, however, narrow this definition to only observe explicit examples of the expressive discourse function in text.

We decompose the general notion of emotions that are conveyed to be valenced reactions, as either a positive or negative polarity based label. There is little consensus as to the set of emotions that humans exhibit, however methods have been put forward to extend these polarities into the realm of emotions (Ortony et al., 1988; Smith & Lee, 2012), so there is the potential for future work to extend this.

The components of expressive discourse when explicitly expressed are often trivial to iden-

tify. Utterances beginning with the personal pronoun *I* followed by an emotive verb often pertain to the expressive discourse function being utilised if they are succeeded by an additional emotion bearing component. Much research in sentiment analysis has observed the expressive discourse function (Mullen & Collier, 2004; Bloom et al., 2007; Dermouche et al., 2013).

3.9.2 Persuasive

Persuasion attempts to perform one or more of the following three actions: to change a decoder's belief or beliefs, to gain a change in a decoder's attitude, and to cause the decoder to perform a set of actions (Miller, 2002).

Sentiment can be viewed as a key component in persuasion, yet it is no trivial feat to define what a positive persuasive utterance is. We define what we shall call contextual and non-contextual persuasive utterances. First, let us observe the non-contextual persuasive utterances. An example of a positive persuasive utterance is: *You should give him a pay rise*. Taking this utterance alone, it is clear that the encoder of the signal is attempting to persuade the decoder to give someone more money for their work, which can be understood to be attempting to elicit a positive action from the decoder, for the benefit of the target of the utterance. However, despite being positively-valenced for the person the utterer is referring to, the utterer may be articulating this viewpoint with a general annoyance around the fact that the person has not received a pay rise. In such a case, there are conflicting sentiments at play between utterer and target, and hence there is a division between the contextual or non-contextual sentiment of an utterance, and the mood of the utterer. In this work, sentiment is considered in reference to the evaluation of an entity in a text, although it is acknowledged that sentiment can be understood from the mood of the utterer.

To contrast this, we must demonstrate a non-contextual negative persuasive utterance. For example, take the utterance *Please fire him*. Here the encoder is attempting to stop the intended target of the utterance from working, by persuading the decoder to ensure they cease working, which is typically seen as something negative (at least in Western societies). The utterer again may be happy when generating this utterance, but due to the negative connotations associated

with losing one's job, this utterance is taken to be negative.

We must also consider the class of persuasive utterances that we describe as contextual persuasive utterances. Again, the determined sentiment refers the content of the utterance, and not the underlying attitude of the utterer. An example of such an utterance is: *Please give me a call*. At first glance, this utterance lacks a clear sentiment. However, if we precede this with the sentence *Great work!*, the above persuasive utterance becomes positive. If instead, we precede the initial persuasive utterance with the sentence *You've messed up.*, our seemingly emotionless persuasive utterance becomes negative. This agrees with the view of Hunston (2011), that indicating an attitude towards something is important in socially significant speech acts such as persuasion and argumentation.

Summary

This chapter began with a description of the patient feedback domain in order to highlight the fundamental differences between patient reviews and other types of review used for sentiment classification. This led to a discussion of the structure of patient reviews, and the organisation responses that are adjoined to the patient comments in the NCSD. These appear to highlight the dominant sentiment of the feedback but do so in a more constrained manner, which is appealing for machine learning approaches to sentiment classification. In order to justify that the responses and feedback are suitable sources of data for investigation, in this chapter, we detailed the annotation study and corpus analysis that was carried out on the NCSD.

The annotation process labelled both Type 2 reviews and the relevant organisation responses, and an agreement study found a substantial level of agreement between the sentiment of the reviews and their responses. Results of the data analysis using corpus linguistic techniques highlighted the aspect-based review style of the Type 1 reviews, in comparison to a more patient-focused review in the Type 2 feedback. Positive reviews tend to be explicit in verbalising the sentiment of the reviewer, whereas negative reviews tended to be more cautious in their approach, conveying sentiment in a less-confrontational and polite manner. While this may

seem problematic, the review responses were shown to be a viable source of for determining review sentiment, which we will investigate in the latter chapters of this thesis. Given the suitability of this dataset for the task of sentiment classification, the following chapter uses the annotated data as a resource for training and testing a number of supervised machine learning classifiers.

CHAPTER 4

AUTOMATIC SENTIMENT CLASSIFICATION OF PATIENT FEEDBACK

Introduction

Sentiment analysis has been undertaken in a number of domains, but the applicability of the task using different review types in the clinical domain has not been critically examined from a supervised machine learning perspective. The task, if successful, would enable health providers to aggregate and analyse a multitude of data in a swift and convenient manner, paving the way for automated decision-making systems.

This chapter examines the applicability of supervised machine learning techniques to the classification of sentiment in patient feedback. At the beginning of this thesis a number of research questions were posed that would examine the extent to which the choice of data, model and feature could affect the performance and outcome of sentiment classification. These were posed to determine if one or a collection of the examined methods can be deemed to be more suitable than others when classifying the sentiment of patient reviews in this domain. We find that while trends emerge when experimenting with the different approaches, we are rarely able to reject the null hypothesis that no statistically significant differences can be found between a majority of the methods, although consistently high performing methods are noted.

This chapter is organised as follows: in section 4.1 we discuss the motivation for examining the applicability of machine learning to sentiment classification in the clinical domain. Section

4.2 discusses the methodology that we will undertake to investigate the relevant hypotheses. Section 4.3 describes our implementation using the Weka toolkit and section 4.4 presents and evaluates the results. The results of a misclassification analysis are given in section 4.5, and the chapter concludes in section 4.6 with an investigation into the classification of patient reviews using only their final sentences, which proves competitive.

4.1 Motivation

Traditionally, patient feedback has been submitted through paper forms rather than in a digital way. However, this trend has gradually changed, and websites such as NHS Choices and Patient Opinion provide interactive portals where feedback can be left. Unlike some review sites, where data is not made available for research purposes, through `data.gov.uk`, patient feedback data has been made available to the general public. These sites pose no demographic constraints upon the user, and so a wide subset of the English population are able to leave comments, leading to a linguistically diverse dataset representative of those with different backgrounds. This is an ideal dataset to capture the variation in everyday language use in relation to the topic of the patient experience. It is also a dataset that will guarantee a variation in the expression of sentiment due to the differences in experiences that people encounter when undergoing medical treatment.

The difference between how sentiment is communicated in patient feedback and other domains such as film and product reviews make it challenging to use currently trained models for analysing the sentiment of a patient review. Using a generic sentiment lexicon (De Smedt & Daelemans, 2012) for the task of the sentiment classification of patient feedback resulted in an accuracy of just 56.42% for binary sentiment categorisation into the categories positive and negative. The technique using this lexicon assigns scores to words that are highly indicative of sentiment in a document, such as adjectives and adverbs. The sum of the scores is then calculated based upon the sentiment-bearing terms that are identified in a document. This accuracy is low for a binary sentiment classification task and highlights the fact that relying on a general

purpose lexicon for the task of sentiment classification in a domain that expresses sentiment in quite a specific way, such as in patient feedback, should be avoided. Approaches that adapt a lexicon to a domain could be examined, but such work would form the basis of a whole other strand of research on the lexical tuning of sentiment lexicons for use in the clinical domain, which is not the intended focus of this thesis.

Instead, this thesis focuses on the use of supervised machine learning techniques for the classification of sentiment in patient feedback to categorise data as either positive or negative. Such techniques are able to successfully learn models of sentiment conveyance from in-domain data and have successfully been applied to domains such as film (Pang et al., 2002) and product reviews (Blitzer et al., 2007). However, little work has thoroughly examined the application of supervised machine learning classifiers to learn models of sentiment based upon the different structures of patient feedback to our knowledge, as we discuss in the literature review in Chapter 2. Therefore, the application of a number of supervised machine learning classifiers that have been used in the literature, in combination with a number of different linguistic features and feature weights are tested in order to investigate the most effective approach to the sentiment classification of patient feedback.

4.2 Experiment Methodology

In this section, we will discuss the methodological approach to examining the research questions set out at the start of this thesis. The methodology follows that of Pang et al. (2002) who were one of the first to examine the effects of machine learning classifiers on the task of sentiment classification, and their methodology stands as one that has been replicated throughout the literature. In this, they develop a dataset with uniform class distribution and perform cross-validation to examine the performance of supervised machine learning classifiers for the sentiment classification of film reviews. Given this approach, different aspects of the task of sentiment classification can be examined by varying the data type, the classification model, and the choice of feature, as we will discuss in this section.

The first research question concerns the sub-types of document that a review can hold, and whether one type of review structure from the sub-types is preferable to use for the training of supervised machine learning models that will be applied to the classification of sentiment into the categories of positive and negative in clinical document sets. The NCSD provides an appropriate document set for this purpose. The reviews in the NCSD are divided into several sub-types, which can be used in the experiments that aim to answer this question.

Three similar machine learning experiments were set up to investigate this first question. The first experiment received a document set of Type 1 reviews as input to the 10-fold cross-validation of several supervised machine learning algorithms. The second experiment used the same collection of supervised machine learning algorithms in a 10-fold cross-validation process, except in these runs a document set consisting only of Type 2 reviews was given as input to the experiments. The third experiment examined Type 3 reviews, a combination of the types 1 and 2, for the machine learning experiments. Again, the same set of supervised machine learning algorithms were tested using a 10-fold cross-validation procedure.

In undertaking these experiments we are attempting to reject the null hypothesis that a variety of classifiers perform approximately the same when applied to the different types of review. If the null hypothesis cannot be rejected, then we can conclude that the examined classifiers are able to learn sentiment regardless of document type, and one document type is not more suited to sentiment classification than another. Evaluation metrics for sentiment classification at the document level used in a number of works in the sentiment classification literature (Greaves et al., 2013; Liu, 2012) are implemented in this thesis to evaluate classifier performance. These include the metrics of accuracy, kappa, precision, recall and F_1, each of which is detailed further in section 4.4.1.

To test if the null hypothesis is true, the results of each classifier trained with no feature engineering will be compared; that is only boolean feature weights with no term removal, lowercasing, or stopword removal will be considered in order to perform the analysis. This could be performed across the variety of feature engineering techniques we employ to evaluate RQ3, however, as there are a large combination this would convolute the analysis and may not produce

a reliable comparison.

If the null hypothesis is shown to be false, what should be demonstrated is that one of the three experiment sets, trained and tested on one of the review types, produces better classification performance than the other two. It could also be the case that one of the experiment types exhibits poorer performance than the other two, signifying the possible inappropriateness of the review type for sentiment classification.

To examine the stability of classifier performance highlighted in the null hypothesis the results on the respective datasets will be ranked over each of the experiments, and results will be compared to measure the performance of the review types. The Friedman and Nemenyi tests will be used for this purpose, which we discuss further in section 4.4.

The second research question considers the choice of classifier for sentiment classification in the patient feedback domain. In the literature, a number of different supervised machine learning models have been applied to the problem of sentiment classification. While it would be appealing to test all possible supervised machine learning algorithms, this is beyond the scope of this thesis. However, in answering this research question the intention is to examine some of the most commonly used supervised machine learning algorithms for the task of sentiment classification and discern any differences that may arise in classification performance due to classifier choice. All models tested have theoretical differences in their approach to statistical classification, therefore no assumption is made that all classifiers will perform equally. However, it is expected that one, or a group of classifiers may consistently perform better than the others.

Five classifiers, from the text classification and sentiment analysis literature are chosen for the experiments, each having previously shown to perform competitively, as discussed in the literature review: Naïve Bayes **(NB)**, multinomial Naïve Bayes **(MNB)**, support vector machine **(SVM)**, logistic regression **(LR)**, and random forests **(RF)**. The text in boldface following each classifier name will be used when reporting the results in section 4.4. The research question aims to discern which of these is best suited to the sentiment analysis of patient feedback. A comparison will be made when no feature engineering has been applied to the input document

set; that is only boolean weighting with no term manipulation is applied. Classifier performance for each review type will be evaluated using the accuracy, kappa and F_1 metrics.

This question will initially be restricted to only data from review types 1, 2 and 3 for the experiments, and will be exclusively trained and tested on data of the same type. However, as we shall see, classifier choice is important in answering question RQ4, where classifiers will be trained and tested on review data of a different type.

Research question three examines the effects of feature engineering on the performance of sentiment classification. The literature has shown that significant gains in classifier performance can be made through the use of combinations of different features and weighting schemes. It is therefore of interest to determine if this holds when classifying the sentiment of review documents in the clinical domain.

Following on from research question two, the best-performing classifier that used boolean features only for classification will be used as a baseline to determine what feature engineering methods are able to improve the performance of sentiment classification. This will be examined across the experiments for review types 1, 2 and 3, and also for the cross-type classification experiments. As well as reporting results for the best-performing single classifier, we will also be aware that feature engineering may boost the performance of previously tested classifiers that did not perform the best on the boolean baseline. Any instances where this was the case will be reported.

In their current representation, the text documents in the NCSD were not in a suitable format for use in the machine learning experiments. The documents were required to be converted into suitable document vector representations in order to be considered as input for the training and testing phases of the experiments. A document vector consists of a set of attributes representing word occurrence information from the text strings of a document. Prior to the construction of the document feature vectors, the following preprocessing steps were applied to all the data used in the experiments in this chapter:

- Tokenisation: Only strings consisting of alphabetic characters are taken as features for each document vector; any others are removed.

- Attribute threshold: Feature vectors will consist of a maximum of a thousand attributes. For example, this will be the most common strings following the stop word removal phase, or the most common word-stems following stop word removal and stemming.

Given the completion of the preprocessing, we choose to experiment with four document weighting schemes that have been successfully used in the sentiment analysis literature. Abbreviations in bold-type will be used to report the relevant result for the given feature variation in section 4.4:

- Boolean features **(bool)** whereby the presence or absence of a term in a document is represented as a boolean variable. Word frequency is ignored, and a marker 1 is given for a word frequency of 1 or more, and 0 if not. This is also referred to as a binary weighting (Paltoglou & Thelwall, 2010).

- Word count **(wc)**: The frequency of a word in a document is marked in the document vector. This is sometimes referred to as term frequency in the literature (Paltoglou & Thelwall, 2010) as it also extends to the frequency of words stems.

- Term Frequency - Inverse Document Frequency **(tfidf)**: Calculated as $log(1 + f_{ij}) * f_{ij} * log(N_{Docs}/N_{iDocs})$ where f_{ij} is the frequency of word i in the instance of document j, and N_{Docs} is the total number of documents and N_{iDocs} is the number of documents in which word i appears.

- Normalised document length **(normalise)**: Normalise the word frequencies of a document to take into account the length of a document.

An initial run through of the experiments with unigram features that varied the type of attribute weight did not result in one type of weight being clearly any better than the others when performing sentiment classification of patient feedback. When computing the average rank, the boolean weighting appears to be the most successful, for type one data, TF-IDF for type two data, and normalised word frequency for type three data. However, no significant difference is found between the usage of the different weighting types. In the literature, the boolean

weighting scheme has been successfully used for sentiment analysis (Paltoglou & Thelwall, 2010; Pang et al., 2002) and text classification (Schneider, 2004), but other schemes such as term frequency have also been found to be successful too (McCallum et al., 1998). As the boolean weighting requires minimal computation and hence shorter processing times, this weighting type is chosen when varying the following document attributes:

- Lower case tokens **(lower)**: Case is normalised before adding to the document vector as an attribute.
- Stopword removal **(lower-stop)**: A list of 571 generic stopwords developed by McCallum (1996) are removed from the document before determining the vector attributes. This follows a process of lower-casing the documents.
- Minimum term frequency: This minimum threshold represents the number of times a term must appear per class in order to be included as an attribute in the document's feature vector. In our experiments, this is set to 1 **(bool)**, 5 **(min5)** and 10 **(min10)**. The feature vectors for **min5** and **min10** use a boolean weighting.
- Word stem **(lower-stem)**: Considers a word stem as an attribute in the document's feature vector. Stems are determined using the Lovins Stemmer (Lovins, 1968).

Finally, research question four tackles a novel problem in the sentiment analysis literature. While previous studies have examined the generalizability of classifiers across domain (Bollegala et al., 2011; Engström, 2004), and others have examined the effects of sentiment classification across different genres of document (Kessler et al., 1997), we examine the effects of different document types on sentiment classification within the clinical domain. Types 1 and 2 offer different purposes in their respective types, yet a reader is able to infer a sentiment through what is written, regardless of format. Therefore, we examine whether it is possible to train on Type 1 documents and test on a document set of Type 2 reviews, and vice versa, without degradation to classifier performance.

We hypothesise that training across the document types will lead to a degradation in performance due to the function of the documents differing. Type 1 documents are supposed to

highlight the aspects of the patient's experience that they liked or disliked, and serve as a summary of the sentiment. Type 2 documents, on the other hand, are longer in length and have a larger vocabulary, and give a more descriptive insight into the patient's experience.

These could be shown to correlate with the description of discourse function proposed by Kinneavy (1969), that the Type 1 reviews form part of the *expressive* discourse function, whereas Type 2 reviews are used for a *persuasive* discourse function. There is some degree of overlap between the two categories, as documents of both discourse function are able to convey a sentiment. The difference in discourse function could be associated with the difference in vocabulary, as was examined in the previous chapter, so despite both exhibiting similar traits by conveying a sentiment, training and testing across these document types will probably be to the detriment of a given classifier. If this is the case, then this has wider implications on the practical application of sentiment analysis and the effects of discourse function on classifier performance.

To evaluate this research question, two more experimentation frameworks will be set up: one that trains on type 1 documents and tests on a document set of type 2, and one that trains on type 2 documents and tests the learned model on a document set of type 1 documents. The same five learning algorithms that we applied to the previous experiments will be applied here, along with the choice of features. The baseline will be a boolean feature set as before. We will compare the outcomes of the accuracy, kappa and F_1 performances with the outcomes for experiments tested on the type 1, 2 and 3 data. A lower performance across these metrics will suggest a degradation in performance and a confirmation of the hypothesis.

The number of documents used for these experiments from each subsection of the NCSD were balanced. Relative information detailing the subsections is summarised by Table 4.1. Type 3 experiments used a combination of type 1 and type 2 data, with the overall document labelling being that of the type 2 advice annotation.

To summarise, five overarching experiments will be run. Abbreviations in bold type will be used to denote the results of the particular training and testing combination in section 4.4:

T1: Train on type 1 reviews and test on type 1 reviews.

Corpus	D_N	W	$D_{avglength}$	$W_{uniq.}$
Type 1				
Positive	750	47875	62	4869
Negative	750	50676	67	5411
Type 2				
Positive	750	44527	59	4587
Negative	750	97408	129	7391

Table 4.1: Type 1 & 2 experiment data statistics. D_N is the number of documents, W is the number of words, $D_{avglength}$ is the average document length in words, and $W_{uniq.}$ is the number of unique words for the given data subset.

T2: Train on type 2 reviews and test on type 2 reviews.

T3: Train on type 3 reviews and test on type 3 reviews.

C1: Train on type 1 reviews and test on type 2 reviews.

C2: Train on type 2 reviews and test on type 1 reviews.

For each of the five experiments, five machine learning models will be evaluated: Naive Bayes, multinomial Naive Bayes, support vector machine, random forest and logistic regression. For each of these models, different word vector representations of the review documents with different attribute weighting schemes will be examined. We limit this to the list previously mentioned in this section. There are other word vector representations that could have been combined, however, we limit our experiments to minimalistic representations in order to study their individual effects upon the outcome of sentiment classification.

4.3 Implementation

The experiments in this chapter are developed to investigate the effects of different review formats in the clinical domain on the outcome of machine learning models applied to classify

patient feedback as either conveying a positive or negative sentiment at the document level. The experiments are implemented in the Weka machine learning environment (Hall et al., 2009). Each review document is stored in a separate text file, each of which is then stored in a directory with a relative polarity naming: positive or negative. These are converted to the desired input for Weka, the attribute-relation-format-file (ARFF) through an inbuilt conversion tool. In this document, instances are stored as a string datatype with a nominal sentiment class label. To run the supervised classification algorithms over the data to train and test the models, the string instances must be converted to a word vector that represents the word occurrence information of documents in the NCSD . A StringToWordVector filter is applied to each document instance to complete the conversion procedure. During this process, attribute filtering techniques can be applied in order to examine the effects of different feature representations on the final classification outcome. The string to word vector filtering techniques discussed in section 4.2 are all available through the Weka API.

The results of applying Weka's StringToWordVector filter to the two example documents 'The surgery was terrible' and 'The nurse was great', for example, would be the following word vector representation:

The	*nurse*	*surgery*	*was*	*great*	*terrible*	...	`label`
1	0	1	1	0	1	...	*neg*
1	1	0	1	1	0	...	*pos*
...	...	...	...	...	...	...	...

The elements in the top row of the word vector denote the document attributes. In the above example, the document attributes are unigrams. In our experiments, we also examine the effects of other attribute representations, such as word stems. In the case of such experiments, following the application of the Lovins stemming algorithm (Lovins, 1968), the unigrams would simply be replaced in the header row by the appropriate words stems.

The above example shows the insertion of only two sample document instances into an example word vector, however, if more documents were to be added to the vector, as would probably be the case for a word vector used to train a supervised machine learning model, new

column attributes may be added to the vector depending on whether the attribute in question was not present in the word vector, and new documents would be inserted after the bottom-most row. The potential for additional elements to be added to the word vector are denoted in the above vector through use of the . . . notation. The attributes in the given examples are shown in an order which highlights the potential overlap of words in different documents, and how a word vector represents this for different documents. The ordering in a practical example would not necessarily be as easily comprehended, so we have simplified this for clarity in the above example.

Below the attributes of the header row, the values contained in each cell are numeric and represent an attribute weighting. In the above examples, these are boolean values; a one denotes the presence of an attribute in the original document whereas a zero denotes its absence. However, these numeric values may take on different types of weight, for example, they might represent the frequency of an attribute in a document or the frequency of the attribute normalised by dividing by the length of the document that it appears in.

The last column of the vector denotes the document instance category labelling; for the sentiment classification task at hand, this nominal value may be *pos*, denoting a document conveying a positive sentiment, or *neg*, for a document denoting a negative sentiment. This is a requirement for classification using Weka, whereby the class attribute must be in the final column of the word vector. When presenting a document for classification, this class attribute labelling is left blank initially, and an appropriate labelling is assigned given the application of a trained classification model.

During the experiments, a baseline word vector is constructed for the classifiers in order to compare the effects of other feature combinations on the sentiment classification of patient feedback. The baseline word vector is constructed with as little feature engineering as possible so as to maintain as much of the original language use as possible for comparison. The word vector consists of unigram document attributes, with a boolean weighting scheme, whereby case information is retained. The retention of this information could be viewed as irrelevant to the expression of sentiment in a document, however as the words contribute to the structure

of the document and enable the coherent utterance of sentiment-bearing terms, these are then left in the word vector for as features. The terms that feature in the vector have a minimum frequency of 1, and weightings are not normalised in respect of document length. This is the simplest feature set with which the experiments are approached, and in section 4.4, results will be reported in respect of what is achieved when using this baseline word vector.

Following from this baseline, features are altered so as to examine their potential effects on classification outcome. Each feature listed in the previous section will alter the baseline in one or more specific ways.

4.4 Evaluation

4.4.1 Evaluation metrics

In this section, we define the metrics that can be used to evaluate and compare sentiment classification methods. Evaluation metrics are defined in order to objectively state the relative strengths and weaknesses of a classifier in relation to other classification algorithms. Evaluation is performed on unseen data, whereby unseen means that the test data has not been used to train the classifier, as this would provide a source of bias to the classifier and over-exaggerate its ability to generalise to classify unseen documents. Additionally, constraints may be loosened to allow a document to belong to more than one class, but for this work, documents are restricted to a single class labelling per document.

The metrics that are used for text classifiers are similar to those used for information retrieval (Manning et al., 2008): accuracy, precision, recall, F_1 and error rate. If the classification

		Predicted	
		System positive	System negative
Actual	Gold positive	True Positive (TP)	False Negative (FN)
	Gold negative	False Positive (FP)	True Negative (TN)

Table 4.2: Example confusion matrix.

problem is binary, that is there are only two classes to classify documents into, then we can produce a 2 by 2 confusion matrix that details the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) from the output of classification. Figure 4.2 shows an example confusion matrix that would be produced given an evaluation procedure carried out on test data. Using the confusion matrix, evaluation metrics are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4.4}$$

$$\kappa = \frac{A_o - A_e}{1 - A_e} \tag{4.5}$$

Accuracy is the basic measure of a supervised machine learning model's classification performance. However, datasets may be skewed and contain a large number of documents in one class, meaning any trivially constructed classifier could achieve high accuracy knowing this fact alone. Therefore, observing the precision or recall are preferable in order to discriminate between classification algorithms. The choice of either metric is application dependent, but viewing both is helpful. However, when one of these metrics performs well this often comes at the detriment of the other (Buckland & Gey, 1994). Due to this, the weighted harmonic mean, the F_1 of the two values is often observed. The Kappa statistic κ (Cohen, 1960) factors in the possibility of chance agreement between classifier outcome and the gold-standard labelled test data (Artstein & Poesio, 2008; Carletta, 1996), where A_o is the observed agreement, and A_e is the expected agreement, calculated by summing the joint probability of the gold-standard label and the classification label over all possible labels. Interpretations of the resulting Kappa statistic value are presented in Table 4.3.

When evaluating a classifier, data is split into a training and a test set. For example the

Kappa Statistic	Strength of Agreement
< 0.00	Poor
0.00 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 1.00	Almost Perfect

Table 4.3: Kappa statistic interpretations (Landis & Koch, 1977)

training set may be the first three-quarters of the data, and the test set the final quarter of the data. However, the class distribution of the final quarter may drastically differ to that of the first three-quarters, and so results from these experiments could give a misrepresentation of classifier performance. Therefore we can vary where the split occurs i.e after the second quarter or the third quarter when dividing training and test set. This process is called cross-validation and it is used to minimise bias in classifier evaluation. We run 10-fold cross validated experiments to examine the performance of the given classification methods.

Our work takes inspiration from the work of Greaves et al. (2013), who ran a number of experiments to examine the performance of sentiment classifiers on NHS patient feedback data. Their work makes notable contributions to the field, and so far as the literature is concerned, their work represents the current state of the art as they examine a number of different machine learning models over the course of their experiments. However, no claims are made about the performance of the classification models, or whether any of the examined methods significantly outperform any of the others, and so may be found to be preferable when constructing a classification system for potential future practical purposes. For this reason, we choose to go beyond a basic suggestion about classifier suitability, and through the use of a robust statistical test, we examine whether one or a series of methods may be preferable for the sentiment classification of patient feedback. This requires a test that is able to simultaneously compare a number of variables to make a claim regarding the significance of the examined methods, which will be

discussed in the following section.

In taking inspiration from the work of Greaves et al. (2013), we can also compare our results to those that were previously achieved in their work. While both of our investigations share the same domain, as the data used in their paper was not available, our datasets differ slightly in both size and review type. A distinction is not drawn between review types by Greaves et al. (2013), and the data used is able to examine other assets of patient feedback that we did not have access to, such as cleanliness ratings and overall comment ratings. These differences are inherent due to the time period from which the data resides: our dataset is from 2012 to 2014, while their dataset is from the years of 2008 to 2011. The datset used in their experiments is also significantly larger than our dataset, using 20,214 comments for training and testing classifiers, in comparison to the 3,000 used in our experiments. The best-performing classifier from their work, that we shall use as a baseline for comparison, was the multinomial Naïve Bayes model, achieving an accuracy of 88.6% and F_1 of 0.89. This result was achieved using only one round of cross-validation. To ensure the integrity of our results, we perform 10-fold cross validation for our given classification experiments.

4.4.2 Friedman Test

The Friedman test will be used to examine the significance of the results (Friedman, 1937). Similar to the analysis of variance method for comparing multiple test results (Fisher, 1932), the Friedman test is a non-parametric method that can be used to determine any significant differences between multiple datasets used to classify sentiment with multiple classifiers. Its default mode is to check for equivalence between outcomes, but beyond a critical value, it indicates whether significant differences in classification have occurred. Given k related samples, either review type, classification model or feature choice, these are ranked over N repetitions of the experiment, varying different review types of classification models to yield the rank r_i^j, where r_i^j for example is the rank of the jth of k classification models on the ith of the N datasets. The null hypothesis assumes the related samples to be equivalent, and so the average ranks over the repetitions should be equal. The Friedman statistic T, equivalent to χ^2 with $k-1$ degrees of

freedom is defined as follows:

$$T = \frac{12}{Nk(k+1)} \sum_{j=1}^{k} \left(\sum_{i=1}^{N} r_i^j \right)^2 - 3N(k+1) \tag{4.6}$$

Iman & Davenport (1980) argue that this statistic is too conservative, and propose a more liberal test for significance:

$$F = \frac{(N-1)T}{N(k-1)T} \tag{4.7}$$

which can then be compared with usual tables for the F-distribution at $(k-1)$ and $(k-1)(N-1)$ degrees of freedom. We use this more liberal version that can be compared with the F-distribution to calculate significance. If significant differences are found, a posthoc method can be applied to determine which datasets yielded significantly different classifications outcomes. By implementing the improved Friedman test (Iman & Davenport, 1980), we are able to use the Nemenyi test to compare the datasets for any significant differences in performance.

4.4.3 Nemenyi Test

The Friedman test is unable to determine exactly which variables significantly differ, so a posthoc test enables the discovery of the significantly different variables. The Nemenyi test (Nemenyi, 1963) is used for this purpose. A significant difference is assumed if the representative average ranks differ by at least the critical difference, calculated as follows:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{4.8}$$

where q_α is the critical value to a confidence level α of 0.05, based upon the studentized range statistic divided by $\sqrt{2}$ (Demšar, 2006).

4.4.4 Critical difference diagram

The results from each experiment are tabulated and presented in Appendix B, and summarised in this section through the use of a number of critical difference diagrams. This technique, introduced by Demšar (2006) enables the results of the comparison of multiple variables to be shown in an intuitive and compact way and is suited to the demonstration of the performance of one variable in comparison to a set of other variables. This work has been used in the sentiment classification literature where a number of experiment variables, such as classifiers, are required to be compared with one another (Smailović et al., 2014). In this visualisation, the axis plots the average ranks of the variables being studied, which in our case are the review types, classifiers and features. The axis has been reversed, so the worst rank is one the left of the axis and the best on the right. This diagram joins variables together through the use of a bold horizontal line in order to highlight that there are no significant performance differences between the given variables. Above the axis, the length of the horizontal line represents the critical difference that is used to compare classification performance. Any value that falls outside the critical difference range when classifiers are compared can be assumed to perform significantly differently from the other classifiers. The diagram caption will give the critical difference value (CD) at a confidence level of $\alpha = 0.05$.

To calculate the Friedman and Nemenyi statistics, and to generate the critical difference diagrams, the Orange toolkit (Demšar et al., 2013) is used. All statistics from which the critical difference diagrams are generated, and also from which the baseline results are reported, are given in Appendix B.

4.4.5 Baseline comparison

Before we investigate whether any of the classification models examined in this thesis perform significantly better than another through the use of the Friedman and Nemenyi tests, we will first compare the classification performance to the established baselines. We have selected two baselines for the experiments in this chapter.

The first is the baseline for each classifier with a unigram feature representation and a boolean feature weighting. This is the most basic supervised machine learning setup for sentiment classification (Pang et al., 2002), and hence serves as a suitable baseline for comparison in each of the experiments.

In Table 4.4, classifier accuracy results are given in respect of each classifier's baseline configuration. Results show that given changes to the representation of a document's feature vector, classification accuracy results are able to surpass that of the baseline for all classifiers over all review types. Notable improvements over each classifier's baseline can be seen in particular for type 2 reviews, where the NB classifier accuracy improves by 6.312%, from a baseline accuracy of 69.547% to 75.859% when normalising the feature weights be the document length. The best accuracy result is reported for type 2 reviews using the MNB classifier, with an increase of 1.348% from 82.722% to 84.070%. Examining the features that gave rise to the maximum classification accuracies, six came through normalisation of the term-weights by the document length, and eleven came through the use of lowercasing the unigrams in combination with a stemming procedure.

Table 4.4: Baseline accuracy comparison for each classifier over each review type. Italicised values are the baseline. The line below each of these gives the maximum accuracy given feature alterations. The string in the brackets denotes the given feature configuration that yielded the improvement. Statistically significant improvements over the baseline are indicated by a ∘ , and are calculated using the paired T-Test ($\alpha = 0.05$).

ML Model	Type 1	Type 2	Type 3	T1 to T2	T2 to T1
NB - baseline	*81.383*	*69.547*	*67.370*	*79.412*	*55.540*
NB	82.336 (lower)	75.859 (normalise) ∘	73.394 (lower-stop) ∘	80.944 (lower)	65.213 (normalise) ∘
MNB - baseline	*81.745*	*82.722*	*82.029*	*83.33*	*70.209*
MNB	82.927 (normalise) ∘	84.070 (lower)	83.542 (normalise)	84.436 (lower-stem)	79.382 (lower-stem) ∘
SVM - baseline	*77.114*	*77.702*	*78.299*	*64.645*	*69.255*
SVM	80.655 (normalise) ∘	78.923 (lower-stem)	78.689 (lower-stem)	70.588 (lower-stem) ∘	75.658 (lower-stem) ∘
RF - baseline	*82.018*	*78.800*	*81.063*	*75.858*	*68.565*
RF	82.109 (min10)	79.781 (lower-stem)	83.359 (tfidf)	76.532 (lower)	73.202 (lower-stem) ∘
LR - baseline	*78.884*	*79.778*	*80.177*	*74.203*	*68.856*
LR	80.473 (lower-stem)	81.560 (lower-stem)	81.585 (normalise)	75.919 (lower)	70.840 (lower-stem)

The second baseline is a comparison to the state-of-the-art work on sentiment classification of patient feedback by Greaves et al. (2013). The best-performing classifier from their work was the multinomial Naïve Bayes model using lower-cased unigram features. They did not discuss what term weighting scheme was used. In their work, relative to our evaluation metrics, only the accuracy and F_1 results are reported for their experiments, as 88.6% and 0.89 respectively (precision of figures given as reported). Table 4.5 shows how this compares to the best-performing configuration of each classifier examined in this chapter.

Table 4.5: Comparison of each classifier to the best-performing classifier of Greaves et al. (2013)

Classifier	Type	Configuration	Accuracy (%)	F_1
MNB (Greaves et al., 2013)	N/A	lower	88.6	0.89
NB	1	lower	82.336	0.834
MNB	1- 2	lower-stem	84.436	0.853
SVM	1	normalised	80.655	0.823
RF	1	lower-stem	82.064	0.828
LR	1	lower-stem	80.473	0.815

In comparison to the work of Greaves et al. (2013), no classification configuration examined in this chapter surpasses the results that they achieved. Despite similar model configurations, the only notable difference between the experiments discussed in this chapter and theirs is the amount of data used for the experiments. They used a dataset of 20, 214 comments for their experiments. In comparison, our dataset only spanned to 1,500 comments per type. From this, we can conclude that if we had access to more annotated data for experimentation, then perhaps results could have matched or surpassed those achieved in their work.

Figure 4.1: RQ1 review type rank accuracy performance comparison (CD = 1.482)

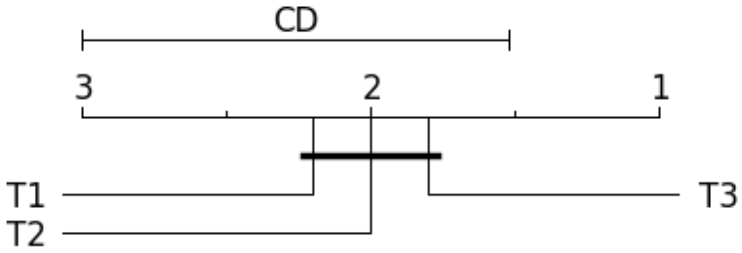


Figure 4.2: RQ1 review type rank Kappa performance comparison (CD = 1.482)

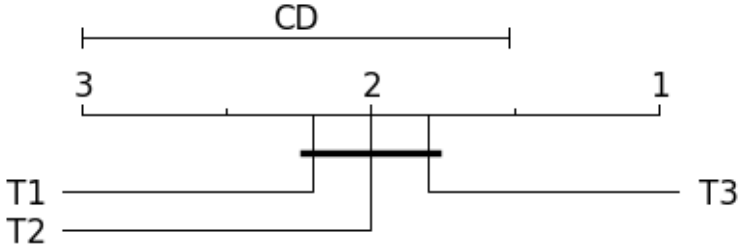
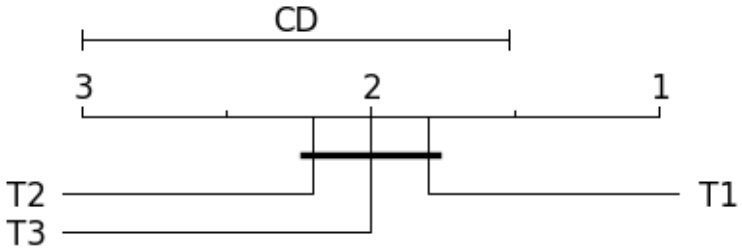


Figure 4.3: RQ1 review type rank F_1 performance comparison (CD = 1.482)



4.4.6 Review type

The first research question investigated the extent to which the type of review affects the performance of supervised machine learning classifiers trained for the task of sentiment classification in the clinical domain. Overall, the mean classification accuracy for the review types was 78.575%. By individual review type, the mean accuracies were as follows: $Acc_{T1} = 80.229\%, Acc_{T2} = 77.710\%$ and $Acc_{T3} = 77.788\%$. Despite the mean accuracy of T1 being 1.653% higher than the overall mean accuracy, the average ranks shown in Figure 4.1 show that the data type with the highest average rank was T3. Calculating significance using the corrected Friedman test showed that there was not a significant difference between the classification accuracies ($F_{2,8} = 0.166$, ns). The overall mean Kappa value was 0.571, and by type, the means were

$\kappa_{T1} = 0.604, \kappa_{T2=0.554}$ and $\kappa_{T3} = 0.558$. The NB classifier produced a difference of 0.277 between T1 and T3. Again, calculating significance using the corrected Friedman test showed that there was not a significant different between Kappa values ($F_{2,8} = 0.166$, ns). Finally observing the F_1 scores, the overall mean F_1 was 0.793. Individually $F1_{T1} = 0.817, F1_{T2} = 0.783$ and $F1_{T3} = 0.779$. The highest average ranking review type was T1, but there was no significant difference between F_1 scores for different review types ($F_{2,8} = 0.166$, ns).

Results demonstrated that no review type is more suited or preferable for the classification of sentiment in text than another when a boolean document representation is used, so the null hypothesis cannot be rejected given the data. We can conclude from this that given a choice of review types, there is not a preferable type to use for sentiment classification, and other factors may be more influential to the outcome of sentiment analysis than review type alone.

4.4.7 Classifier choice

Figure 4.4: RQ2 classifier rank accuracy performance comparison (CD = 3.522)

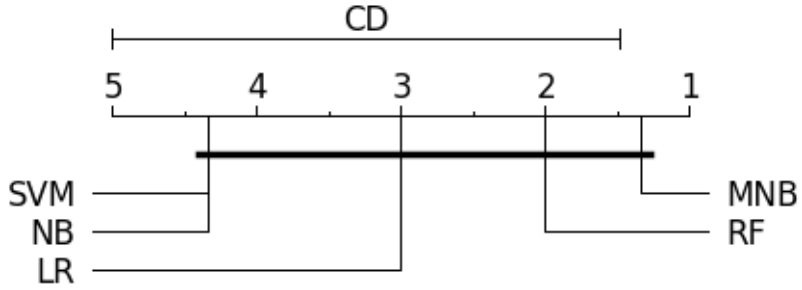
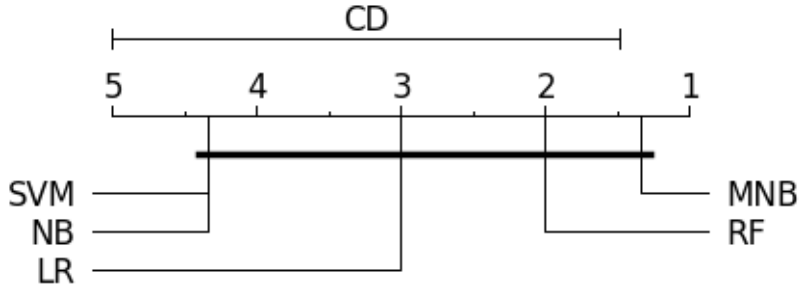
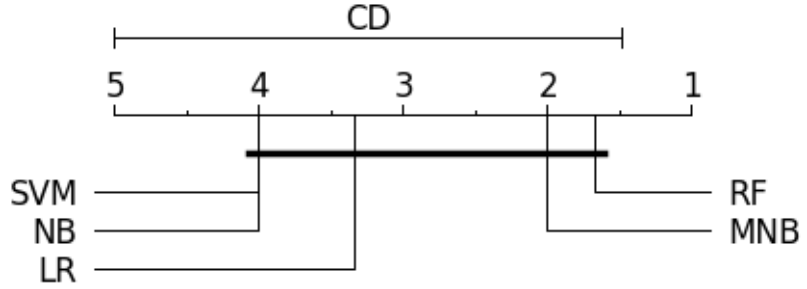


Figure 4.5: RQ2 classifier type rank Kappa performance comparison (CD = 3.522)



The second research question examined the extent to which the classifier choice affected

Figure 4.6: RQ2 classifier type rank F_1 performance comparison (CD = 3.522)



the performance of supervised machine learning classifiers trained for the task of sentiment classification in the clinical domain. The overall classification performance means were $Acc = 78.576\%$, $\kappa = 0.571$ and $F_1 = 0.793$. The best classification accuracy was achieved with the MNB classifier on the Type 2 reviews ($Acc_{T2} = 82.722\%$), and the worst was achieved with the NB classification model ($Acc_{T3} = 67.370\%$). These results were mimicked by the Kappa and F_1 statistics. The corrected Friedman test showed that there was not a significant difference between the classifier outcomes across all data types ($F_{4,8} = 5.499$, $p < 0.05$). The MNB and RF classification algorithms consistently rank higher than the other methods, but do not perform significantly better than any others for the given data sets. We can therefore conclude that for the examined classifiers, the choice of algorithm will not significantly affect the outcome of sentiment classification, although there is a tendency for the MNB and RF outperform the other classifiers.

4.4.8 Choice of feature representation

The third research question examines the performance of the features and weightings individually over the three review types, and results from the classification of the documents using different features for each of the three review types is reported. The overall mean accuracy across data types is 78.533%, overall mean Kappa is 0.571, and overall mean F_1 is 0.791. T1 produced the greatest mean accuracy ($Acc_{T1} = 79.224\%$) and within this, the lowercase boolean feature produced the greatest mean accuracy ($Acc_{lower} = 80.882\%$). TF-IDF is the best feature weight for T2 reviews ($Acc_{tfidf} = 78.568\%$) and for T3 was normalised score

term-weighting ($Acc_{normalise} = 78.813\%$). When observing the Kappa statistic, the lowercase feature is the highest ranked for T1 ($\kappa_{lower} = 0.618$), for T2 the lower-stem is the best feature ($\kappa_{lower-stem} = 0.572$), and for T3 the normalised boolean weight feature is the best-performing feature ($\kappa_{normalise} = 0.596$). The F_1 follows a similar trend, with the lowercase features performing the best for T1 ($F1_{lower} = 0.822$), TF-IDF for T2 ($F1_{tfidf} = 0.796$) and normalised feature weighting for T3 ($F1_{normalise} = 0.802$).

We are only able to reject the null hypothesis when observing the F_1 measure for the T1 reviews ($F_{8,32} = 2.380, p < 0.05$). In this case, the lowercase boolean feature is highlighted as performing significantly better than a lower-case with stop words removed feature vector. However, no other significant differences can be determined. With the remainder of the experiments that explore the effect of feature selection upon the classification of sentiment in the clinical domain, average ranks are consistent for some features as previously shown. This would suggest that of the tested features, one will not significantly improve the performance of the supervised machine learning models trained for sentiment classification over a standard binary feature representation in this domain. However, some features such as lowercasing and normalised weightings may be more useful than others on particular review types.

Table 4.6: Key for feature abbreviations in the critical difference diagrams for the feature choice experiments.

Abbreviation	Detail
bool	Boolean term weighting.
lower	Lower case strings with boolean weighting.
lower-stem	Lower case word stem strings with boolean weighting.
lower-stop	Lower case strings with stopwords removed with boolean weighting.
min5	A minimum term frequency of 5 with boolean weighting.
min10	A minimum term frequency of 10 with boolean weighting.
normalise	Attribute weighting normalised relative to document length.
tfidf	Term frequency - inverse document frequency term weighting.
wc	Word count term weighting.

T1

Figure 4.7: RQ3 feature choice rank (T1) accuracy performance comparison (CD = 5.372)

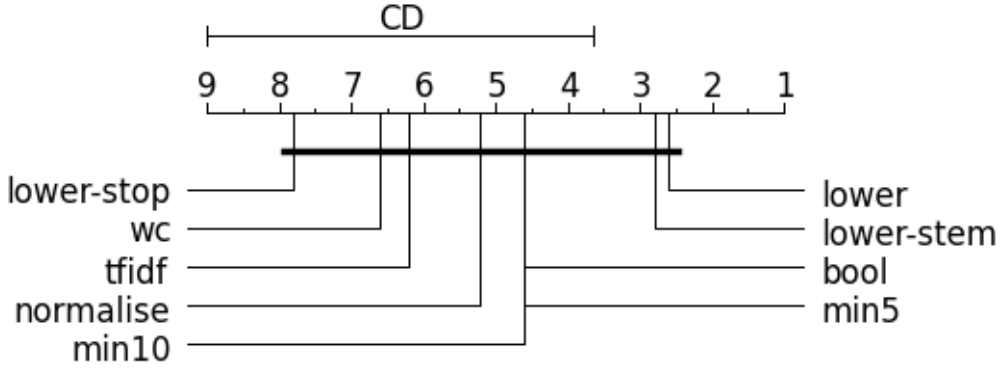


Figure 4.8: RQ3 feature choice rank (T1) Kappa performance comparison (CD = 5.372)

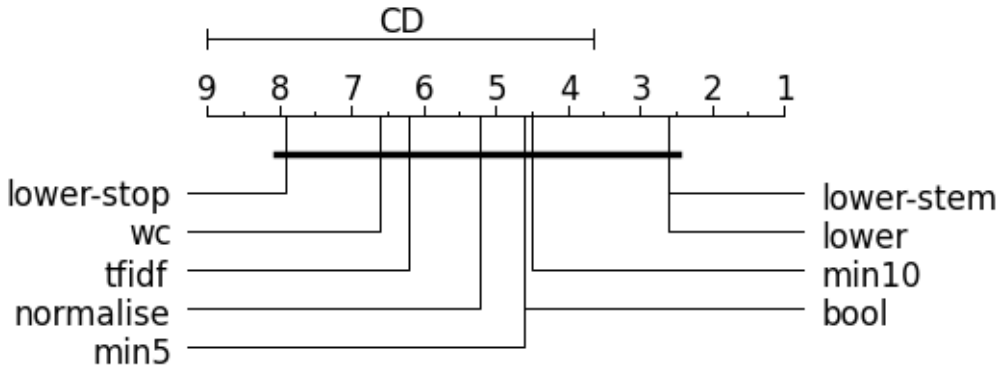
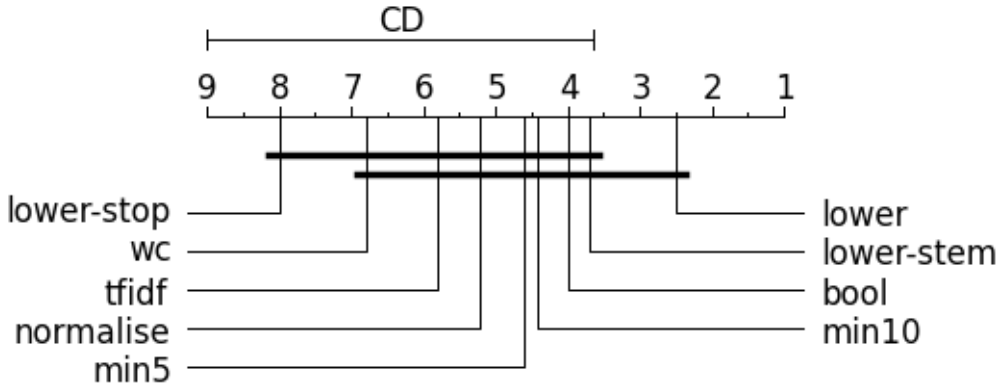


Figure 4.9: RQ3 feature choice rank (T1) F_1 performance comparison (CD = 5.372)



T2

Figure 4.10: RQ3 feature choice rank (T2) accuracy performance comparison (CD = 5.372)

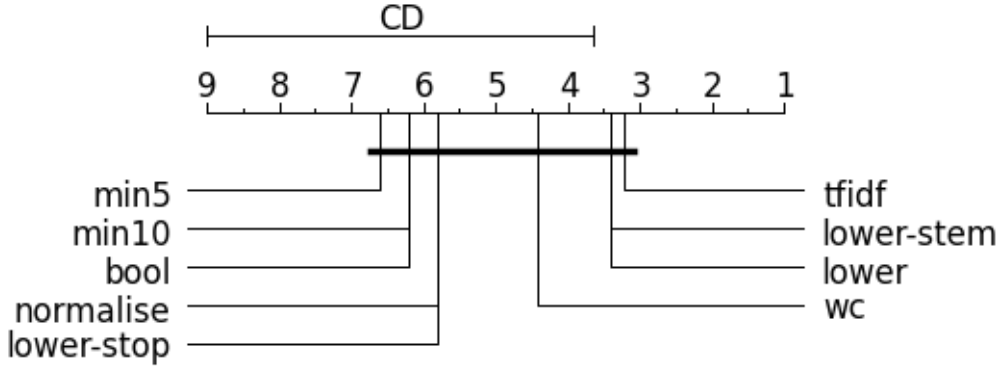


Figure 4.11: RQ3 feature choice rank (T2) Kappa performance comparison (CD = 5.372)

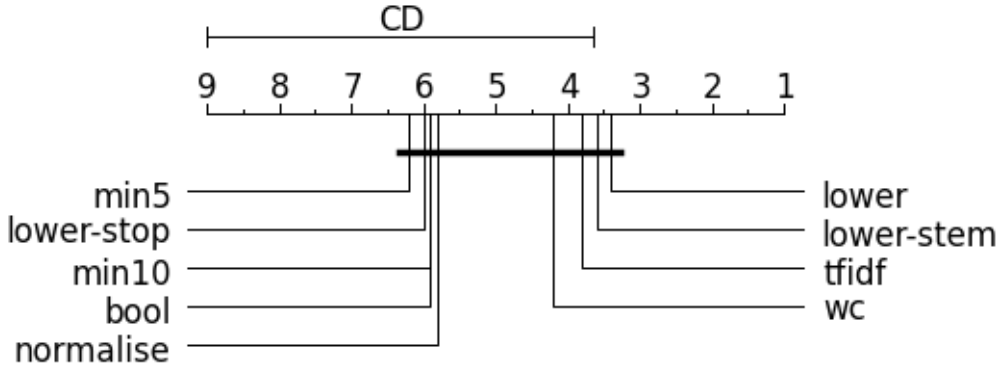
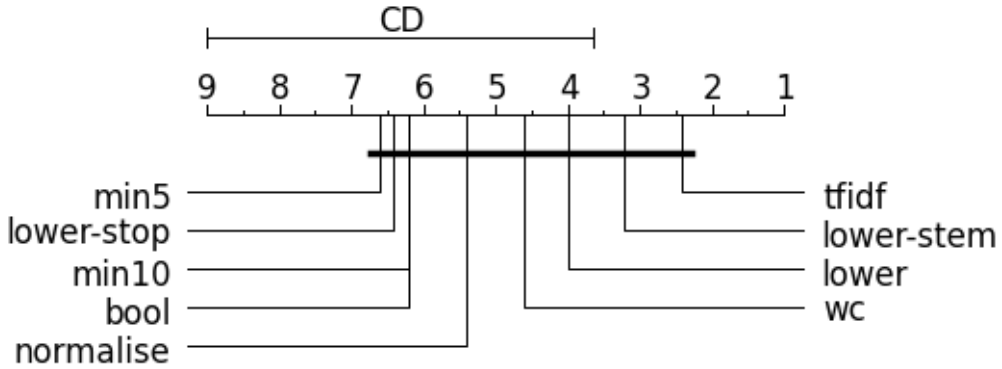


Figure 4.12: RQ3 feature choice rank (T2) F_1 performance comparison (CD = 5.372)



T3

Figure 4.13: RQ3 feature choice rank (T3) accuracy performance comparison (CD = 5.372)

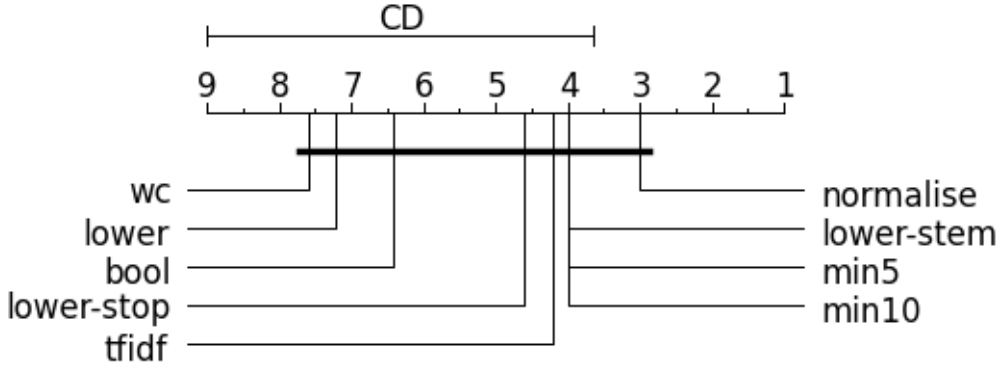


Figure 4.14: RQ3 feature choice rank (T3) Kappa performance comparison (CD = 5.372)

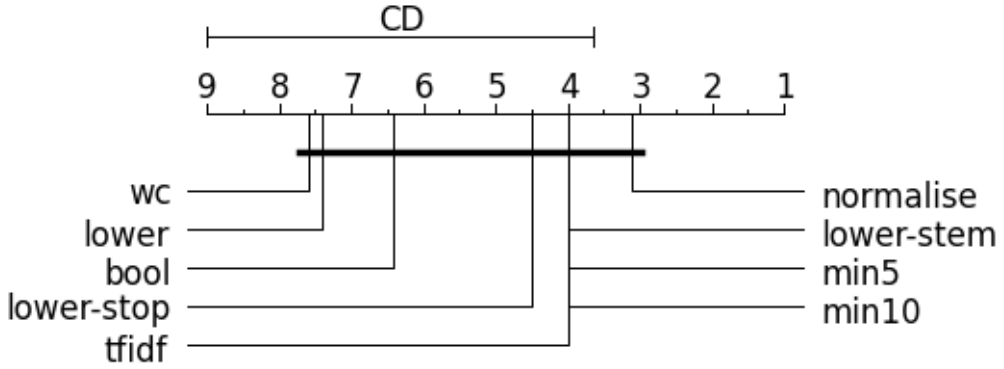
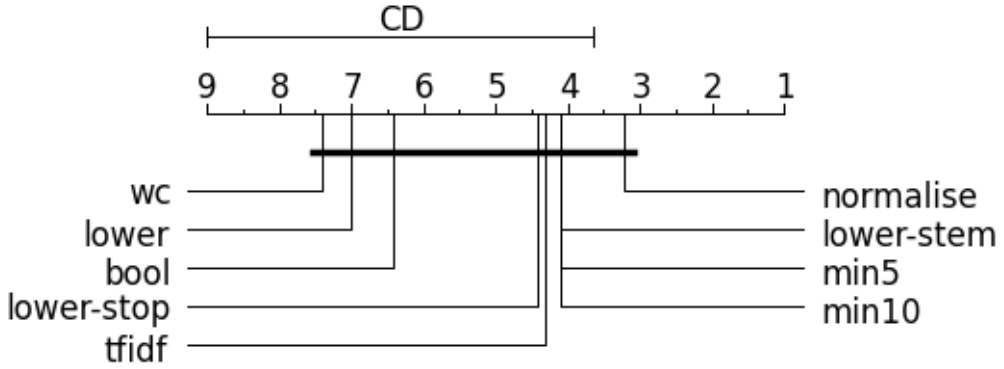


Figure 4.15: RQ3 feature choice rank (T3) F_1 performance comparison (CD = 5.372)



4.4.9 Cross-discourse results

Figure 4.16: RQ4 review type rank accuracy performance comparison (CD = 2.728)

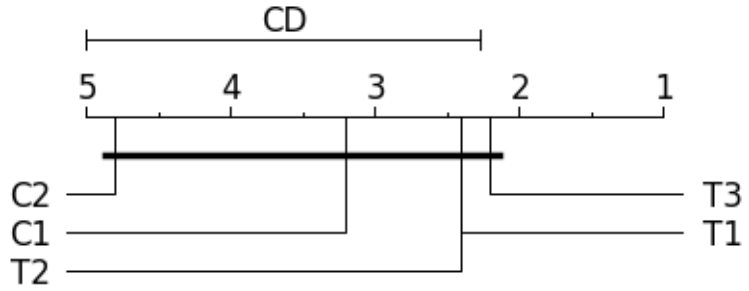


Figure 4.17: RQ4 review type rank Kappa performance comparison (CD = 2.728)

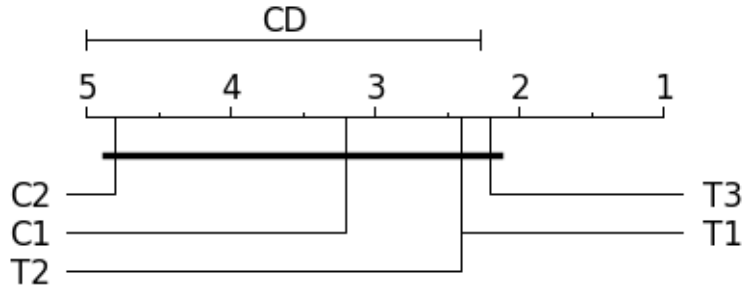
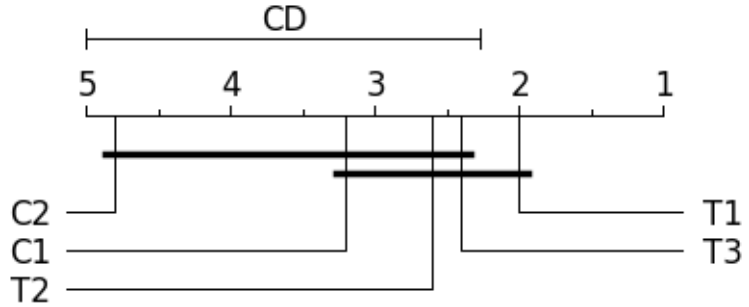


Figure 4.18: RQ4 review type rank F_1 performance comparison (CD = 2.728)



The final research question concerns the extent to which training on a document with a different review type to the review that the model will be tested on affects the outcome of sentiment classification. We evaluate this hypothesis by comparing the cross-type experiments with the within type experiments discussed in section 4.4.6. The overall mean accuracy for the five classifiers over the five types of review experiment is 75.540%, and individually the mean accuracies of C1 and C2 are $Acc_{C1} = 75.490\%$ and $Acc_{C2} = 66.485\%$ respectively. The overall

mean Kappa is 0.510 and F_1 is 0.754. The Kappa statistic is especially poor for all classifiers experimenting on the C2 data, with $\kappa_{NB} = 0.113$.

The data is not sufficient to reject the null hypothesis in this case when observing accuracy ($F_{4,16} = 3.463, p < 0.05$), however when observing the F_1 score of the classifiers over the datasets the model that is trained and tested on Type 1 reviews significantly outperforms the model that was trained with Type 2 data and tested on Type 1 data ($F_{4,16} = 3.692, p < 0.05$) . No significant differences can be detected amongst the other data types with the data that were examined. This significant difference would suggest that Type 2 data may not be as well suited to sentiment classification as originally imagined. However, as Figures 4.16 and 4.17 show, the Type 3 review structure, the combined Type 1 and Type 2 obtains an average rank slightly above that of Types 1 and 2, however again the difference is not significant.

4.5 Misclassification Analysis

While the classification performance was adequate, in comparison with other text classification experiments the results were far from perfect. A large number of experiments were performed in the previous sections, and to evaluate the misclassifications of all would lead to much confusion. Therefore, we focus on the misclassifications of an MNB classifier with boolean features on Type 2 reviews, one of the best-performing experiments.

In total, of the 260 misclassifications that were made, 92 positive reviews were incorrectly classified as negative, and 168 negatives reviews were classified as positive. The average length of the negative misclassified files was 42.60 (min = 1, max = 393, SD = 52.352) words, and the average length of the positive misclassified files was significantly more at an average of 128.75 (min = 5, max = 825, SD = 144.221) words.

Two problems are exposed in the misclassification analysis: the probabilities of terms calculated in the training phase and the effect of spelling errors. Take the following verbatim comment, whose correct labelling is negative:

"very bad experience Simple communicaton"

In Table 4.7, the attributes are assigned the following probabilities when training the MNB classifier:

Table 4.7: Probabilities of a word given the class associated with the input "very bad experience Simple communicaton"

Attribute	Negative	Positive
bad	0.001	7.762E-4
experience	0.002	0.003
simple	2.113E-4	1.350E-4
very	0.004	0.008
communication	5.284E-4	1.345E-4

The first error in classification is exposed with the words *experience* and *very*. Given the usage of these terms in the training data, these are given a higher association value with the positive class than the negative class, resulting in the misclassification of the document.

The second error in the document is the incorrectly spelt word *communicaton*. Spelling correction is not applied to the input, but if it were, an appropriate weighting could have tipped the balance of classification into the correct negative class, and hence helped label the document appropriately.

An examination of the word *good* in the Type 2 reviews indicates another potential reason for misclassification. Nine reviews that contain the word *good* from the negative reviews were incorrectly labelled as positive. *Good* can be assumed to be a positive word, but in the misclassifications, five of the instances of the word are preceded by a linguistic construct that negates the positive sentiment of *good*. Such constructs are difficult to encode in supervised machine learning classifiers. Attempts have been made to incorporate the negating function into the machine learning process by altering the features to indicate that these were under the influence of having been negated, for example, by changing prefixing all affected terms with the prefix `NOT-` (Pang et al., 2002). However, these features were found to have negligible effects on the overall outcome of classification, despite the major shift in sentiment that they indicate.

The problems found in this misclassification analysis are innate to supervised machine learning and are non-trivial problems to tackle. However, there is the potential that a process that could consider the classification of a related document may be of use in reconsidering the given prediction made by a classification model where there may be ambiguity in the overall labelling of a document, for example, where spelling errors or sentiment shifters were present when in the review instance that was being classified. We will discuss this in the following chapter.

4.6 Sentiment classification using final sentences

Type 2 reviews are substantially longer than Type 1 reviews. In turn, the reviewer's sentiment that translates to the overall document sentiment can get lost in a mixture of linguistic constructs, such as negations and contrasting phrases that form the core of a review. However, what becomes apparent from analysis of the review text is that the final sentence tends to summarise the reviewer's overall sentiment. This is a summary of a document and is quite typical of natural discourse (Goldstein et al., 1999).

Becker & Aharonson (2010) undertake two psycholinguistic experiments to demonstrate that the most salient part of a review for computing overall document sentiment is the final sentence. While the distribution of review labellings of participants between whole reviews and last sentences show similarities, the average latency time for labelling given the final sentence was found to be significantly shorter than when examining other isolated sentences of a text, which Becker & Aharonson conclude implies a computational efficiency for review polarity given the final sentence only. While this indicates psycholinguistic feasibility, this is based on an underlying human comprehension, which is not necessarily mimicked in a computational system. However, the computational studies that have examined the use of final sentences for sentiment classification tend to support this result when considering the role of sentiment in different domains. For example, Beineke et al. (2003) examine the use of positional information to generate sentiment summaries. They find that the final 5% of a review enables the generation of particularly useful summaries due to the richness of sentiment-bearing terms. Similarly, Pang

& Lee (2004) find that the last sentence of a review is more indicative of document sentiment than an introductory sentence and therefore is a useful feature that can be used for the polarity classification of movie reviews using supervised machine learning models. In experiments using both an NB and SVM classifier, use of the final sentence only in classification as opposed to the first sentence yielded significant improvements in classification accuracy, from approximately 59% to 66% for the NB, and 56% to 64% when using the SVM. They note that this behaviour is to be expected when classifying film reviews as authors tend to describe the plot in the first part of a review and tend to conclude with their clearly stated opinions. We believe that this could be the general case for reviews, and expand this notion also to patient feedback, whereby patients describe their treatment in the first part of a type 2 review and conclude with how they felt the procedures went, clearly notifying the reader whether a good or bad experience was had. Mukras (2009) find that in two review corpora on actors and motor vehicles, there is a consistent sentiment richness, a notion based on the frequency of sentiment-bearing terms, in the latter part of a review. Despite this consistency, in the actor corpus, there is also richness near the beginning, and the vehicle corpus in the main body. It is unclear of the polarity of the sentiment richness, and whether there is an overlap or discussion of different aspects that the reviewer has different opinions towards. Finally, Biyani et al. (2013) examine the use of last sentence only features in the polarity classification of posts in a cancer survivors network. They identify unique posts types, which either offer direct or indirect emotional support in the forums and show that extracting features for classification from the final sentences yields increases, albeit minimal, in classification precision and recall.

Table 4.8 gives examples of sentences that begin and conclude positive and negative reviews from the type 2 sub-corpus. The length of each review is variable, with an average of 23.460 words per final sentence and 24.783 per first sentence. These lengths may seem unnaturally high for a single review sentence, however, the identification of a sentence in a review can be error-prone due to shortcomings in tokenisers or because of sentences in a given text failing to be concluded with a full stop. Some documents, for instance, insisted on using semi-colon's to end sentences, which would be ignored by a sentence tokeniser, or lack a space between the

full-stop and the proceeding start of a new sentence which again would be ignored.

Given reviews that have no associated star rating, we question whether classifying only the final sentence of a review is a better feature type to use as opposed to the document as a whole. Doing so would result in a compressed document representation that was still representative of the overall document sentiment while requiring a smaller feature vector. We experiment with the best-performing classifier from the Type 2 reviews, the MNB classifier, with the best-performing feature representation: the lower cased, boolean feature representation with a maximum feature vector size of 1,000 terms per class. This experiment set up resulted in the following when using the full length reviews: accuracy = 84.073%, κ = 0.681, precision = 0.884, recall = 0.818 and F_1 = 0.842. This will be used as a baseline for this experiment set.

Given the results of Pang & Lee (2004), we hypothesise that classification using final sentences only will be competitive with classification when using a complete document. Based on the aforementioned work we also hypothesise that final sentence only classification will yield better results than classification using the first sentence of a review only. As review length is variable, we do not examine other single sentences from a document review.

The results shown in Table 4.9 for first and final sentence classification are not better than those achieved when considering a complete document in calculating a feature vector for classification. However, these are competitive, and highlight that while more detail may lead to increased classification accuracy, kappa and overall F_1, a competitive result can be achieved from a document of patient feedback a fraction of the size. Considering the average document length for a type 2 review is 96.122 words, a document length some 76% shorter in length is able to produce results within 4% accuracy and .05 F_1 when the final sentences are used.

Table 4.8: Sample first and final sentences of type 2 reviews

Sentiment	First Sentence	Final Sentence
Positive	I took myself into A & E via taxi one morning three months ago with severe stomach pains.	My life was saved.
	I waited two weeks to see a Breast Specialist.	A very clean hospital displaying a friendly atmosphere.
Negative	I was initially referred to the NHNN Autonomic Unit in Easter 2010.	I understand that it is a busy hospital but I think this is appalling and an unacceptable length of time to be left without any proper treatment.
	I went to A& E after a visit to the fracture clinic for a replacement of a lost thumb splint which is required for my broken thumb.	Did not represent the NHS in a good light whatsoever.

Table 4.9: Results of classification when using the whole review, the first sentence, and the final sentence only with the MNB classifier.

	Whole Review	First Sentence	Final Sentence
Accuracy (%)	84.073	76.409	80.110
κ	0.681	0.524	0.597
Precision	0.884	0.765	0.807
Recall	0.818	0.764	0.801
F_1	0.842	0.764	0.804

Summary

This chapter evaluated the ability for sentiment classification to be undertaken on a collection of patient feedback documents with varying review structures. Given the NCSD, we examined four research questions, each pertaining to an aspect of sentiment classification. Experiments related to the research questions were developed to determine the structure of review data that is best suited to sentiment classification, the supervised learning model that is best suited to the task, the best feature type for the task, and whether cross-type learning was a possibility in this domain. Our evaluation ranked the variables and highlighted potential trends in the results, but experiments using the Friedman test and the post-hoc Nemenyi test rarely rejected the null hypothesis that a particular model, feature or review type produced significantly better results than any other. We can take this as a positive result, despite no significant model being found. The choice of model, feature and data structure can be left in the hands of the user, and they may instead prefer to experiment with the trade-off between model performance and computational efficiency instead.

CHAPTER 5

SENTIMENT CLASSIFICATION IN CONTEXT

Introduction

In natural language processing, the context of a text can be used to help clarify its meaning. By definition, the context of a text immediately precedes or follows it, where in the scope of this thesis a text is defined as a word or a document. As both sentiment and meaning are interlinked, determining and suitably understanding the context of a text is an important step in calculating its sentiment.

In this chapter, we review the role of context in the task of sentiment analysis and how this has been examined in the literature. We first demonstrate the context-sensitive nature in which sentiment is conveyed at the document level and how this poses problems for a system that is developed for the task of sentiment classification. We demonstrate that this task is in fact highly context dependent and similar to other open problems in natural language processing, such as word sense disambiguation. Given this relationship, we question whether methods for word sense disambiguation could be applied to the polarity disambiguation task of document-level sentiment classification. Knowledge-based and corpus-based approaches are examined, which leads us to conclude that a hybrid approach that operates at the document level may be suitable for our purposes.

5.1 Rules for opinion identification

Put simply, the task of identifying opinions in a text can be approached by formalising a set of rules. In the computational sentiment analysis literature, such rules have been developed by Ding et al. (2008) and Liu (2010), which were expanded in further work by Liu (2012). The proposed sets of rules consider the compositional semantics of lexical items that lead to a particular sentiment being conveyed and in turn can be used to identify a given sentiment. These rules operate at a level of abstraction above a base lexical level and require suitable knowledge of word combinations to operate correctly. These rules are concise, yet cannot model the intricacies of language that limit the classification of sentiment in text. The rules for simple sentiment identification are presented in Backus-Naur Form:

$$\langle \text{POSITIVE} \rangle \models \langle \text{P} \rangle \tag{5.1}$$

$$\langle \text{POSITIVE} \rangle \models \langle \text{PO} \rangle \tag{5.2}$$

$$\langle \text{POSITIVE} \rangle \models \textit{sentiment-shifter}\ \langle \text{N} \rangle \mid \langle \text{N} \rangle\ \textit{sentiment-shifter} \tag{5.3}$$

$$\langle \text{POSITIVE} \rangle \models \textit{sentiment-shifter}\ \langle \text{NE} \rangle \mid \langle \text{NE} \rangle\ \textit{sentiment-shifter} \tag{5.4}$$

$$\langle \text{NEGATIVE} \rangle \models \langle \text{N} \rangle \tag{5.5}$$

$$\langle \text{NEGATIVE} \rangle \models \langle \text{NE} \rangle \tag{5.6}$$

$$\langle \text{NEGATIVE} \rangle \models \textit{sentiment-shifter}\ \langle \text{P} \rangle \mid \langle \text{P} \rangle\ \textit{sentiment-shifter} \tag{5.7}$$

$$\langle \text{NEGATIVE} \rangle \models \textit{sentiment-shifter}\ \langle \text{PO} \rangle \mid \langle \text{PO} \rangle\ \textit{sentiment-shifter} \tag{5.8}$$

The above eight rules form the basis for a sentiment classification system. The productions P and N denote an atomic positive and an atomic negative expression, whereby, as will be expanded in later rules, the expressions may be a word or phrase. PO and NE are positive and negative sentiment-bearing expressions that are composed of multiple expressions in combination. The sentiment shifters stated perform the role of inverting the sentiment expression they appear alongside. For example, a sentiment shifter in combination with N or NE will result

in a positive sentiment. Liu (2012) explains that the shift function is an abstract notion that may appear lexically before or after a sentiment expression, but does not include this in the rules, so here we have expanded the rules appropriately to incorporate this. Polanyi & Zaenen (2006) give examples of sentiment shifters for English, whereby negators, such as *not*, or modal auxiliaries, such as *could* or *should*, flip the scope of the sentiment being expressed.

The positive and negative atomic expressions are defined as follows:

$$\langle \mathrm{P} \rangle \models \textit{positive-sentiment-word-or-phrase} \tag{5.9}$$

$$\langle \mathrm{P} \rangle \models \textit{desirable-fact} \tag{5.10}$$

$$\langle \mathrm{P} \rangle \models \textit{within the-desired-value-range} \tag{5.11}$$

$$\langle \mathrm{P} \rangle \models \textit{produce a-large-quantity-or-more resource} \tag{5.12}$$

$$\langle \mathrm{P} \rangle \models \textit{produce no,-little-or-less waste} \tag{5.13}$$

$$\langle \mathrm{P} \rangle \models \textit{consume no,-little-or-less resource} \tag{5.14}$$

$$\langle \mathrm{P} \rangle \models \textit{consume a-large-quantity-of-or-more waste} \tag{5.15}$$

$$\langle \mathrm{N} \rangle \models \textit{negative-sentiment-word-or-phrase} \tag{5.16}$$

$$\langle \mathrm{N} \rangle \models \textit{undesirable-fact} \tag{5.17}$$

$$\langle \mathrm{N} \rangle \models \textit{deviate-from the-desired-value-range} \tag{5.18}$$

$$\langle \mathrm{N} \rangle \models \textit{produce no,-little-or-less resource} \tag{5.19}$$

$$\langle \mathrm{N} \rangle \models \textit{produce some-or-more waste} \tag{5.20}$$

$$\langle \mathrm{N} \rangle \models \textit{consume a-large-quantity-of-or-more resource} \tag{5.21}$$

$$\langle \mathrm{N} \rangle \models \textit{consume no,-little-or-less waste} \tag{5.22}$$

Rules 6.9 to 6.22 demonstrate the formation of positive and negative atomic expressions. These can loosely be grouped into the categories of words that: (a) convey sentiment, (b) exhibit

desirability of fact or a value range, and (c) the business-leaning rules concerning production and consumption of what is deemed to be a resource or waste. At the core of most sentiment classification systems are the rules 6.9 and 6.16, which are used to identify the sentiment of a recognised atomic expression in a text. Despite the general form of the rule, differences between system performances occur when the knowledge source that contains the sentiment-bearing words and phrases differ.

Difficulties start to arise when considering how to develop the remainder of the rules given above as what is a desirable or undesirable objective statement is unintentionally and somewhat ironically, subjective. For example, the statement: *the nurses worked quickly* may imply a positive opinion about the nurses, but also may be interpreted as a negative statement depending on the quality of care that the surrounding context of the statement implies.

The identification of resources and waste, and actions that consume or produce these requires specific in-domain construction, often from a domain expert. Such rules may be highly precise, but the payload of constructing such rules may not be consistent with the time required to create them given their nature.

The rules for combined sentiment expressions and potential sentiment items are defined as follows:

$$\langle \text{PO} \rangle \models \textit{decreasing} \langle \text{N} \rangle \mid \textit{increasing} \langle \text{P} \rangle \tag{5.23}$$

$$\langle \text{NE} \rangle \models \textit{decreasing} \langle \text{P} \rangle \mid \textit{increasing} \langle \text{N} \rangle \tag{5.24}$$

$$\langle \text{PO} \rangle \models \textit{no-decreased-quantity-of} \langle \text{NPI} \rangle \mid \textit{increased-quantity-of} \langle \text{PPI} \rangle \tag{5.25}$$

$$\langle \text{NE} \rangle \models \textit{no-decreased-quantity-of} \langle \text{PPI} \rangle \mid \textit{increased-quantity-of} \langle \text{NPI} \rangle \tag{5.26}$$

$$\langle \text{NPI} \rangle \models \textit{negative-potential-item} \tag{5.27}$$

$$\langle \text{PPI} \rangle \models \textit{positive-potential-item} \tag{5.28}$$

PO and NE are defined as sentimental expressions that are composed of many constituent expressions that include either an increasing or decreasing atomic sentiment expression (6.23-

6.24) or an increased or decreased quantity of a potential sentiment-bearing item (6.25-6.28). An example of an NPI or PPI is a noun, such as "cost" or "waiting time" that alone convey no sentiment, but given the context of a preceding positive or negative adjective, they enable the production of a sentimental expression.

5.1.1 Difficulties

It seems reasonable to assume that the application of all of the aforementioned rules in union should lead to a functional sentiment classification system. However, Liu openly admits a central flaw: these rules, in particular, the atomic sentiment expressions, may appear in potentially innumerable lexical forms and so are difficult to list and therefore recognise. If a computational system is unable to recognise the lexical form of a rule then the sentiment of a text cannot be appropriately determined. Given this conclusion, a system with a robust knowledge base with a full coverage, constructed with the given rules in mind, should enable a sound and functional computational approach to sentiment analysis to be developed.

However, it is difficult to anticipate the way in which a knowledge source may adequately handle conflicts in the rule usage. For example, given the utterance *'more doctors are needed'* a distinct lack of a resource is implied, and therefore a negative evaluation of the situation should be attached to the utterance. Given the aforementioned rules, if we were to apply rule (6.23) $PO \models increasing\langle P\rangle$, in combination with rule (6.9) or (6.28) whereby the term *doctor* is a member of a subset of positive or potentially positive words, then an overall positive sentiment would be deduced by this system, which is an incorrect labelling in the case of this example. This highlights just one of the shortcomings in a rule-based approach to sentiment classification whereby category polarity assignments may inadvertently cause problems if not handled appropriately.

As well as this, figurative uses of language in a text being classified by the given rules would be problematic. Sarcasm is a problem throughout sentiment analysis that is detrimental to classification outcomes (Riloff et al., 2013). It is a linguistic phenomenon that relies on the flaunting of common communicative expectations, so what appears positive on the surface and

could be identified so by the opinion identification rules is in fact intended as negative, and what seems to be a negative document is in fact intended to be positive. Liu (2012) finds that sarcastic utterances are present but relatively infrequent in online reviews, however, in online debates, it is a much more common feature and requires appropriate handling. In tweet data, Maynard & Greenwood (2014) find that in a small corpus of sarcastic tweets, 75% of the data refers to extra-contextual sarcastic elements that require additional knowledge sources to the tweet alone in order to identify the tweet polarity.

The two problems that we have discussed in this section: the flexible sentiment categories in text and the handling of figurative language, are currently difficulties that are purely computational. We are able to handle these occurrences with little difficulty, yet find that incorporating our cognitive processes into a system to handle such phenomena is relatively complex. The key to unlocking the relative sentiment under these conditions would appear to be understanding the context with which an utterance containing either or both of these elements is given. Therefore when attempting to handle both of these problems computationally, we should consider how contextual features can be used to help automatically determine the sentiment of a document.

5.2 Polarity disambiguation

The task of disambiguation in natural language processing, in particular, word-sense disambiguation (WSD) is a major, unsolved challenge. It affects the outcome of tasks, such as parsing, information retrieval and relative to this thesis, sentiment classification. These tasks are all related to the problem of word sense disambiguation as a meaning or sense of a word can often differ given its usage within different contexts.

Included in the sense of a word, among other aspects of meaning, is the prior polarity that is associated with a given word. As discussed previously, the polarity of words in a review document contribute to the overall polarity of the review, but occasionally the compositional rules required to identify the polarity of an expression do not operate as intended, leading to misclassified documents as was shown in the previous section. The problem of malfunctioning

rules can be blamed on the shifting polarity of some words that when used in a particular context change its overall polarity, which links this task back to that of word-sense disambiguation.

General approaches that have been proposed for the task of word-sense disambiguation can be grouped into two major approaches: knowledge-based and corpus-based (Montoyo et al., 2005). Each defines and uses context in a different way.

Knowledge-based approaches to word sense disambiguation use machine-readable dictionaries to give context to the words that are to be disambiguated. Methods that make use of the knowledge-based approach attempt to count the overlap between dictionary definitions and the context words of the given node word (Lesk, 1986), the construction of word-sense vectors (Wilks, 1990) and the use of semantic relations provided by resources such as WordNet to disambiguate a given word (Mihalcea & Moldovan, 1999). These methods, while robust, do make the assumption that the knowledge source has been appropriately annotated and also has appropriate coverage of the particular domain that it is to be used in.

Corpus-based approaches use machine learning methods to learn the context that a word appears in within a corpus and train a classifier given this knowledge. As with machine learning approaches to sentiment classification, this relies on suitably annotated sense-data in order to develop the models.

The aforementioned methods operate on the word level, as this can be seen as the level at which meaning is conveyed, so meaning can be determined by examining the individual words of a document. In combination, however, words produce a document-level meaning. Attempting to generate a knowledge source for all possible combinations of words and documents that acts as a document level lexicon is infeasible due to the potential combinations of words given a grammar of a language. This makes the notion of a document *sense* somewhat unrealistic. Despite this, the overall polarity of a document can be determined, as is shown by the rich back catalogue of work on review classification (Pang & Lee, 2008). Following from this there is a possibility that context can be used to guide the calculation of the polarity of a document, and there is the potential that word-sense disambiguation methods may be able to be adapted to document-level polarity annotation. We explore this further in the following sections of this

chapter.

5.3 Sentiment in context

In this section, the various approaches to using context in the task of sentiment classification will be outlined and discussed. Returning to the idea of knowledge sources in sentiment classification, a sentiment lexicon is a list of words with a prior positive or negative sentiment labelling. When using this in a system that matches the words from the lexicon to words in a document, the sentiment of the words in the document is influenced by the surrounding context, whereby in a particular context the word may have the opposite sentiment to that listed in the lexicon, or no sentiment at all (Wilson et al., 2005b).

This has a profound effect on the construction of resources for sentiment classification. The phenomena of contextual sentiment was first examined by (Hatzivassiloglou & McKeown, 1997) when attempting to computationally determine the semantic orientation of adjectives. Their work focused on the concept of sentiment consistency in text, whereby given a seed set of adjectives with known polarity any adjectives conjoined to the seeds in the text by a conjunction such as *and* should communicate the same polarity as the seed. This concept was used to expand a seed set into a knowledge source for sentiment analysis. Kanayama & Nasukawa (2006) discuss the benefits of sentiment consistency, but note that the context window of only conjoined terms used in the previous work is limited. Therefore, they examine both intra-sentential sentiment consistency and inter-sentential sentiment consistency that adjacent sentences offer. The inter-sentential approach is found to be competitive with the intra-sentential approach for developing a lexicon of polar terms. Ding et al. (2008) further examine the challenge posed by words that are both domain and context dependent, and find that attempting to tackle the domain specificity of words is insufficient alone for sentiment classification considering intra and inter-sentential context, as terms such as *long* and *small* require contextual information in order to classify correctly. Inter-sentence information is also used to build probabilistic models that can classify sentiment (Sadamitsu & Yamamoto, 2008). Lexical cohesion based approaches are

also used to expand the scope of terms in a graph in order to give additional context to individual terms (Devitt & Ahmad, 2007).

Building on methods proposed for corpus-based WSD, Akkaya et al. (2009) adapt these methods to show that disambiguating the word sense of subjective terms is an important element of polarity classification. Using a combination of subjective annotated data and an SVM classifier, subjective word sense disambiguation is shown to be more successful than WSD, implying the applicability of the machine learning approach to polarity disambiguation.

Pang & Lee (2004) proposed a graph-based approach for polarity classification that operates on the intra-sentential level. In this approach, they first extracted the subjective sentences from a review. This was first carried out by training an NB classifier to detect subjective portions of a document. This proved fruitful for polarity classification, with results improving on using the whole review text as features for classification. This initial approach assumed that sentences operate in isolation from one another, despite intuitions indicating otherwise.

To incorporate context and improve the performance of the method, the degree of proximity between sentences was calculated by implementing the constraint that nearby sentences will share the subjective or objective properties of its neighbours. Given a threshold T, specifying the maximum distance any two sentences can be separated by, a distance function $f(d)$ is then applied that determines the degree of influence a sentence has in respect of potential decreasing proximity. The distance functions $f(d) = 1, e^{(1-d)}$, and $1/d^2$ are all experimented with, where the best accuracy score is taken given the application of each individual function, but no claims are made as to the optimal function in these experiments. A weighting factor c where $c \in [0, 1]$ at intervals of 0.1 is additionally applied in order to control the degree of influence. A higher value of c makes it unlikely that sentences in close proximity will be classified differently.

In order to incorporate the contextual information of nearby sentences, the sentences of a review were transformed into a graph-based model. From this, a minimum-cut formulation was applied to the document graph to split the document into subjective and objective sentences. Sentences from the document were represented as the graph's nodes and edge weights between the nodes denoted a level of association between the sentences. A cut of the graph would parti-

tion the nodes of the graph into two distinct sets, and these sets would have an associated cost value based upon the sum of the edge values crossed when partitioning the graph. A minimum cut for the graph is then defined as that which yields the minimum cost when partitioning the graph into two distinct sets of nodes. Further details of the method are given in their paper, but relevant to this thesis is a non-increasing function that was calculated to detect the influence of a nearby sentence in respect of the potential decreasing proximity. In this function, an association score between sentences was calculated where the distance between two sentences was considered to be proximal if it was less than a threshold T, where $T \in 1, 2, 3$. This yielded encouraging results for classifying reviews by polarity, but further improvements were returned when incorporating contextual information through a minimum-cut method applied in order to incorporate the contextual information of nearby sentences.

This model was shown to be effective for sentiment classification and was adapted by several other works that attempt to tackle the problem (Wilson & Wiebe, 2005). Due to the data with which these experiments have been undertaken, and their review based instances, these researchers have naturally only used the minimum-cut formulation between sentences in a document to classify a document by an overall stance that is taken. Sentiment, however, does not stop at the utterance boundary.

5.4 Inter-document context

The context provided by the words and sentences in a document are not the only points of reference that can be used when computationally determining the sentiment of a text. In the previous examples, words were observed to give the contexts of other words and likewise, sentences were used to determine the context of other sentences. However, the sentiment labelling process for reviews tends to be applied at the document level, whereby the polarity expressed by the document is summarised by a positive or negative labelling. The question that is therefore posed and examined in this section is whether the sentiment of one document can be used as a given context for another document.

Often documents that are given as input to sentiment classification systems appear in isolation. This is particularly relative to individual review documents that are used as test beds for the development of sentiment classification algorithms. By *in isolation*, we mean that the documents are written without a clear relationship to another document. For example, an online review is a standalone document that does not necessarily need to be produced as a part of a dialogue or discourse. Given this, the meaning should be clear from the content of the document on its own. However, a system to determine the polarity of a document may still struggle if a poor knowledge source that has not encountered certain lexical items in training, or where contextual word and sentence features within the document are insufficient to determine the overall review sentiment or are possibly misleading.

Taking online reviews as standalone documents may have been a sound assumption to make for earlier works in the field of sentiment classification, such as those by Pang & Lee (2004); however, the way in which online reviews are now submitted and interacted with has evolved. Now, given an online review, a chain of comments may follow the original review. These responses are indicative to the reviewer that their review has been acknowledged, and that others agree or disagree with their opinion. For example, the following exchange of comments may occur on an online review site for a GP's surgery between the reviewer A and the respondee B:

A: *Considering the long queues of people, thank you for the time you gave me.*

B: *Thank you for your comment. It is always nice to get good feedback and we do try our best to offer the best possible service we can.*

A supervised machine-learning classifier may struggle when attempting to assign a sentiment category label to the patient A's review, as the review presents contrasting sentiments. Here, the queue in the utterance is described as 'long', which in this context has negative connotations due to the long wait that is implied. Following this phrase, the expressed sentiment shifts, with the patient now giving thanks for the service provided to her. The main sentiment of the text is defined by the latter segment of the utterance in this example. Computational approaches have been proposed to address the nature of contextual valence shifters, but these are

typically rule-based, and therefore non-trivial to augment. Unless prior data is suitably annotated, and an example features in the training data, which in a unique case such as this may be unlikely, a supervised machine-learning approach to this may not adequately capture the shift in sentiment.

While some machine learning approaches may struggle to accurately classify the patient's review when it is treated as an isolated document, the response clearly indicates the polarity of the review. An inter-document context is presented here whereby through the presence of the response, a sentiment can be inferred about the original review. We can liken this effect to a sentence level phenomenon described by Somasundaran (2010): *"If two sentences are close to each other in time, it is likely that they belong to the same discourse context, thereby increasing the likelihood that the opinions in them are related"*. No assumptions are made as to the interlocutors involved in the discourse, only that there is a likelihood of overlap in the discourse context of their utterances. While this operates at the sentence level, we believe that this assumption can also be expanded to the document level, under the presence of particular relationships and constraints.

5.4.1 Constraints

As shown in section 6.1, high-level rules for opinion identification provide a strong theoretical basis for opinion identification in text but can be inaccurate in practice due to the lack of a context. Due to this, both Asher et al. (2008) and Somasundaran (2010) have developed an unintentionally complimentary linguistic schema that considers rules for identifying opinion in text given a set of discourse constraints.

Somasundaran defines relations between the targets of opinions and the relationship between opinion bearing utterances. Opinions are related through identification of opinion frames, compromising of three item tuples: these contain the polarities of the two utterances and the target of opinion that connects them. An interlocutor may present a reinforcing opinion or a non-reinforcing opinion in what they say of write, each of which contributes to the resulting polarity of the utterance. In the proposed framework, rules for calculating an overall stance given the

presence of two polarity bearing expressions and a relationship between the opinions can be either reinforcing or non-reinforcing:

1. Reinforcing discourse-level opinion relations:

   (a) $< +, +, same >$

   (b) $< -, -, same >$

   (c) $< +, -, alt >$

   (d) $< -, +, alt >$

2. Non-reinforcing discourse-level opinion relations:

   (a) $< +, -, same >$

   (b) $< -, +, same >$

   (c) $< +, +, alt >$

   (d) $< -, -, alt >$

These rules capture the expression of opinion at the discourse level. A *same* target relation indicates that the targets of opinion are related, while *alternative* is indicative of the fact that the opinions relate to different entities in the discourse. Combinations of these rules lead to an overall stance being revealed. The stance is clear for rules 1(a) and 1(b), however in the reinforcing relations of 1(c) and 1(d), it is unclear of the overall sentiment that will be assigned in light of the different focus in topic. Similarly, for the non-reinforcing relations, rules 2(a) and 2(b) appear to be weighing up the pros and cons of an opinion target, while 2(c) and 2(d) seem to indicate that positive or negative aspects of different opinion targets are being mentioned in a discourse.

Asher et al. (2008) in a similar vein investigate the effects of rhetorical relations on opinion expression. An opinion can belong to one of the categories of either *advise, judgement, sentiment*, or *reporting*. Each is subdivided into further subcategories of expression.

Rhetorical relations between opinions belong to one of five categories: *correction, contrast, support, result* and *continuation*. These relations describe how various opinion expressions are combined into either a strengthening (support, continuation, result) or weakening (contrast, correction) relationship for opinion expression. However, the proposed scheme is only applicable to an utterance from a single speaker or author, and is not developed with generalisation across instances where there may be more than one author or speaker in communication with each other.

These approaches lend themselves well to sentiment classification in a discourse, as in turn taking during a discourse, an interlocutor will either reinforce his own opinions or refute the opinions of others, which can be modelled given the proposed relations of the two schemes. Somasundaran's approach enables the polarity to be tracked over numerous utterances, limited to a ten sentence window. The work considers multiple discourse participants but does not attempt to determine a global polarity of the discourse.

Constraining the context of a discourse to only the utterance of a single interlocutor is justified by the assumption that sentences belonging to the same speaker are more likely to be related through the notion of an opinion frame than sentences that belong to different interlocutors in a discourse. However, given the constraint between different discourse participants that they were involved in discussing the same topic, then it may be justifiable to assume that these constraints could also be used across the participants for determining the global sentiment.

5.4.2 Relationships

The relationships between groups of documents must be considered in order to determine their suitability in yielding a context. For example, Agrawal et al. (2003) study the behaviour of individuals in an online newsgroup forum to examine whether they support or oppose a discussion topic. They find that a characteristic behaviour of those in the newsgroup is that users more often reply to a post that they disagree with. Using this heuristic, they apply a partitioning algorithm to a user interaction graph to group those exhibiting similar stances on an issue. In doing so, they argue that the agreeing or disagreeing link behaviour is more important than the text

of each post, which contradicts the typical assumptions of work in automated text classification and sentiment analysis. Results from their experiments showed significant improvements when only considering the links of a network, as opposed to classifying the stance of a document using its text as the main feature of classification. They demonstrate that both the NB and SVM supervised classification models perform poorly. However, this could be attributed to the size of the dataset used to train the classifiers, which only takes documents from approximately fifty authors, and no details are given about the number of tokens or document lengths. Similar experiments dispute this claim (Murakami & Raymond, 2010) and highlight the relevance of the text of an utterance in classifying stance on a similar dataset, but their approach uses a rule-based classification system as opposed to the machine learning models used by Agrawal et al. (2003).

The general position of users in online debates can be computed by examining the local information, the remarks between one another, in relation to a central topic. Murakami & Raymond (2010) use two heuristics to compute a user's stance in a debate: (1) users either comment directly about the central topic of the debate, stating whether they support or oppose the motion and (2) users respond to other users remarks thereby through local agreement or disagreement with other users they indirectly support or oppose a proposed topic. As agreement can be voiced in both a direct or indirect manner, this makes computing a user's stance difficult when analysing the surface syntactic form of the remarks for sentiment classification.

To compute a user's stance, Murakami and Raymond (ibid.) consider the content of a remark alongside a graph of the utterance structures in relation to one another in a debate. The first step they use to infer support or opposition is to compute the degree of disagreement between any two users by taking the link structure and text of their adjacent replies and using this to calculate link weight between nodes, the users, in a network. The second step is to apply a maximum cut method to the network in order to split the users into two disjoint sets based upon the maximum sum of the links. This is computed under the assumption that when the disagreement is higher, link weighting will be higher and two groups of users with opposing positions will naturally emerge.

The degree of disagreement is calculated by finding a reaction coefficient. This is defined as a function of two users $r(i, j)$. Given this, a comment and its reply are assigned a local position label: $agree, disagree, neutral$. The reaction coefficient between participants is then defined as:

$$r(i,j) = \alpha N_{disagree}(i,j) + \beta N_{neutral}(i,j) + \gamma N_{agree}(i,j)$$

where $N_{opinion}(i, j)$ represents the number of remark pairings with a particular opinion as the local label between the participants i and j, and α, β and γ are predefined weights, where $\alpha = 1, 0.5 \geq \beta > 0$ and $-1 < \gamma < 0$. To calculate the reaction coefficient between pairs, a rule-based agreement classifier alongside a machine-translation based sentiment analysis system is applied. Results of opinion classification showed moderate levels of precision but low recall.

In summary, the work of Murakami & Raymond (2010) shows that the proposed method of observing the local positions of users as opposed to looking at the global scale and aggregating user's opinions is beneficial to overall stance classification where there are multiple participants. Their results show significant improvements over that of Agrawal et al. (2003).

Despite this, their work is not without potential drawbacks. Despite a reply being directed at a particular remark, it often contained opinions about the main topic. While not directly stating this, these could possibly be in disagreement with what the labelling process finds to be the label. For example, a user may disagree with a previous poster due to the irrelevance of a remark, yet still, support the overall position.

However, what this paper reinforces is that when a user is faced with a series of comments on a website, they will typically respond to the comment that is most relevant to them. Here the content of the original comment is important in determining the overall sentiment of the response. Based upon the relationship information, can the focus of the problem be reversed and can the response be used to identify the sentiment of the original comment? In the work on forums and newsgroup interaction, without knowing the context of the reply, this would appear to be difficult. However, in certain situations, the response is limited in such a way that it can contain vital context that indicates the overall sentiment of users' interactions.

5.4.3 User Interactions

Using a social network as a basis for experimentation, Leskovec et al. (2010) examine the positive and negative connections that form between users. Here, the theories of balance and status are applied to users in a social network. *Balance theory*, formulated originally by Heider (1946), explores the notions of signed edges in triangular relationships between three individuals. The signed edges of a triangle represent either a positive or negative relationship between individuals. In combination in a triangle with three positively signed edges, this represents the notion that a friend of my friend is also my friend, and if there are two negatively signed edges, and one positive, a friend of my enemy is also my enemy.

Status theory, introduced by Leskovec et al. (2010), is based on the observation that a signed link from A to B could imply a status hierarchy between A and B. If there is a positively signed link from A to B then this could mean that A thinks that B has a higher status than they do. Conversely, a negative link from A to B could mean that A believes that B has a lower status than they do. These views are not limited by level and can be propagated through a directed graph.

These two theories, however, produce contradictory links when there is a positive link between A and B and B and C and one is trying to predict the polarity of the sign between A and C. Balance theory dictates that there should be a positive link, whereas status theory would create a negative link.

5.4.4 Reciprocation

Given the user interactions, Leskovec et al. (2010) examine directed links between two users, based on the presence of a dialogue between the two; referred to in their work as a reciprocal link. The dialogue may have conveyed the same sentiment from both participants, or differing sentiments from each participant in the dialogue. They find that balance theory can be widely observed when there are examples of link reciprocation. However, approximately 4% of the edges studied reciprocate an existing link. Table 5.1 outlines edge reciprocation statistics from

Epinions	Count	Probability
$P(+\|+)$	38,415	0.969
$P(-\|+)$	1,204	0.031
$P(+\|-)$	1,192	0.692
$P(-\|-)$	560	0.308
Wikipedia	Count	Probability
$P(+\|+)$	2,509	0.945
$P(-\|+)$	145	0.055
$P(+\|-)$	193	0.706
$P(-\|-)$	80	0.294

Table 5.1: Table from Leskovec et al. (2010) detailing edge reciprocation statistics. The probability $P(X|Y)$ gives the probability of edge X reciprocating edge Y.

Epinions and Wikipedia.

In Table 5.1, the probability of a positive edge being reciprocated is found to be at least 0.945. The probability of a negative edge being reciprocated is somewhat lower, 0.294 for the Wikipedia network and 0.308 for Epinions. The interesting finding is that given a negatively signed initial edge, the responding edge has a relatively high probability of being positive. For Epinions, this is 0.692 and 0.706 for Wikipedia, which is shown to be indicative of the status theory coming to fruition.

This work introduces the concept of polarity flowing through a network in a concise formalism, but it does not consider how user utterances may be used in expressing a particular signed opinion about another user.

West et al. (2014) extend the ideas of Leskovec et al. (2010) by developing a model that predicts an individual's opinion of another from the signed social network that they are embedded in by considering user utterances alongside network structure. Using the same theories from social psychology of social balance and social status, triangular networks in a graph are identified that are used to predict the polarity of opinion of one user in respect of another. The Wikipedia administrator request data is used for experimentation, alongside a corpus of Congressional de-

bates (Thomas et al., 2006). The sentiment signal is strong within the Wikipedia dataset but is difficult to interpret in the Congressional data due to noise and a lack of directionality. If there were some way to introduce a specific directionality to the problem and to reduce the noise, then this would enable a more accurate computation of the expressed sentiment in the data.

Miller et al. (2011) investigates the phenomenon of sentiment flow in blog posts, and posits the question: "does the sentiment of one post influence the sentiment of its immediate neighbours?"

In their method, the sentiment score for each post is calculated by observing the relevant scores from two sentiment lexicons: SentiWordNet (Baccianella et al., 2010) and the General Inquirer lexicon (Stone, 1966). By labelling the posts, they are able to examine how sentiment cascades through linked posts in a network. When a parent node is objective, the child node exhibits objective qualities, and similarly, when a parent node is subjective, the child is also. As more posts are added to a post thread and the depth increases, findings show that a sentiment tends to become more ingrained and polarised, which supports the notion that sentiments are reinforced from one post author to another.

An alternative analysis of sentiment flow is taken by Nalisnick & Baird (2013), who analyse the sentiment dynamics of dialogue in Shakespeare's Hamlet. For each character participating in an act, the sentiment of the character's dialogue is calculated and assumed to be directed at the character immediately *following* the current speaker. This assumption has limitations, as the topic of sentiment may not necessarily follow from speaker to speaker in a dialogue. However, classification results based on this assumption are encouraging.

Somasundaran et al. (2007) develop a system to perform sentiment classification of the dialogue recorded in a corpus of meeting data. The opinion classification of each turn in the corpus is not solely dependent upon the local features, the text of an utterance, but also on the class labels of related opinions and whether the links are reinforcing or non-reinforcing. Local information is considered in the scope of global links between utterances when classifying the sentiment. In this framework, information flows between sentiment-bearing text spans. By applying a process of iterative classification, the polarity of previously unknown or ambiguous

sentiment-bearing utterances may be labelled using the contextual knowledge of target links. For this to work, a manual annotation process for not only opinions but also the linking of the targets of opinions and the opinion frames. There are clear benefits of taking into account the contextual, neighbourhood information when classifying sentiment. This was based on a graph-based collective classification framework Bilgic et al. (2007).

Mukherjee & Liu (2013) also examine the role of agreement and the reciprocation of sentiment between users in an online forum and responses to online reviews. Through the development of a clustering and ranking technique, relevant multi-word expressions are discovered that generate a lexicon of agreement and disagreement terms, *AD-sentiment expressions*. In combination with an SVM classifier and a range of feature selection techniques to determine the overall interaction between a pair of users as either agreeing or disagreeing. The AD-expressions are found to be the most useful features for classifying an interaction as agreeing or disagreeing, as would be expected in such a task. They yield statistically significant improvements over a baseline that doesn't incorporate the proposed features.

Pair interactions of at least twenty turns are used as the data source for expression extraction in the hope that a rich dataset of AD-sentiment-expressions would be developed. While valid, this seems arbitrary, and discounts shorter interactions that may be just as rich in content. This is also dismissive of more concise dialogues between users, that come to an agreement resolution in a more efficient manner, which could even occur in the smallest possible dialogue, the minimal two turn utterance structure.

In classifying a pair interaction as either agreeing or disagreeing, Mukherjee and Liu (ibid.) concatenate the interactions between authors into a single document and then classify the interactions as a single document. This seems rational, but agreeing and disagreeing, despite the argument put forward that they are extensions of sentiment, in this case, are sentiment independent. One can agree or disagree with someone either willingly or begrudgingly. By willingly agreeing, a positive sentiment will be exchanged between the two authors, while agreeing begrudgingly, despite an underlying agreement on a particular topic, a negative sentiment will be conveyed overall. Due to this, the observation of Mukherjee and Liu (ibid.) that AD-

expressions are an extension of the traditional sentiment classification framework cannot hold, but are weakly correlated linguistic constructions.

The issue of sentiment classification is carefully circumnavigated in this instance. The problem with the sentiment transaction in a dialogue is that it is notoriously implicit at times. However, while much research in sentiment analysis has been undertaken to tackle the ambiguities that sentiment conveyance in natural language poses, few have looked to contextual features of multiple, related documents.

5.5 Responses as Context for Reviews

The literature reviewed so far has highlighted the use of context at the document level. In this section, we will examine further sources of context for sentiment classification.

Mukherjee (2014) introduces the notion of different types of review response or comment expressions (*C-Expressions*) on social networking sites, that he categorises as follows:

1. Thumbs-up - conveying a positive sentiment towards the given review.
2. Thumbs-down - conveying a negative sentiment towards the given review.
3. Question
4. Answer Acknowledgement
5. Agreement
6. Disagreement

Out of the six categories of C-Expression, four intuitively exhibit sentiment-bearing qualities: thumbs-up, thumbs-down, agreement and disagreement. The categories question and answer acknowledgement are not restricted to conveying only a single polarity. Questions tend to be posed in order to gain a clarification, and answer acknowledgement tends to be associated with the positive thanking act. The C-Expressions that are examined are used to generate a

term lexicon, words of which are consequently used as features for comment response classification of a randomly selected 2000 document set using an SVM classifier. C-Expressions in combination with maximum entropy priors are shown to yield the best classification results.

During the process of manual sentiment annotation in this study, C-Expression labels were found to be overlapping in some instances. For example: the sentiment of a Thumbs-Down and Disagreement C-Expression overlapped, and the Thumbs-Up and Acknowledgement and Question did so too. No reported overlap of thumbs up with agreement is cited, which correlates with the findings of Agrawal et al. (2003) that responders mainly reply to commenters that they disagree with.

Mukherjee views these categories as independent to the problem of sentiment classification and so does not attempt to use the labellings of a C-Expression to improve sentiment classification of the original review, although he notes that they are good indicators of the quality of a review. Mukherjee focuses on product reviews as part of his study, and so these categories are limited by the domain. However, service reviews, such as patient feedback, offer one more type of C-Expression that augments the set: the professional response that we discussed in previous chapters.

5.5.1 Responses to Patient Feedback

Given the work of Mukherjee, it becomes apparent that related documents can provide a context for computational approaches to sentiment classification. The NHS Choices website is neither a forum nor a social network; however, it does have an interactive element to it. When posting a review to the site, the reviewer is able to receive a response, visible on the website, which responds to the content of the original review. In the corpus analysis in Chapter 4, we gave an analysis of the distribution of terms in the responses, to which key terms emerged following a keyness analysis. It was apparent that the responses were revealing of a sentiment that was related to the content of the review. Therefore, comment responses on the surface appear to be viable sources of sentiment context with which to classify patient reviews.

Perspective is key to the use of the response in the classification of review sentiment. The

review is written from the perspective of the reviewer, and the response is written from the perspective of the responder. In constructing a relevant response, the author of the response attempts to understand the review from the perspective of the reviewer. Therefore, the response must be constructed in such a way so as to acknowledge the sentiment that the reviewer conveys in their feedback. If this holds to be true, a sentiment analysis system can only benefit from the perspective offered by such a response.

The challenge is then to incorporate these responses into the classification process. One way would be to only classify the response, and then base the classification of the review on the labelling of the response, thereby ignoring the content of the document as carried out in previous work (Agrawal et al., 2003). This method has been shown to be naive (Murakami & Raymond, 2010) however, and incorporating the textual features of both the review and the response into the classification process may be preferable. Unlike other work, we are not considering a social network or multiple postings in a forum, but only a single, professional response to a review. Therefore, graph-based division methods would not be applicable in this instance. A better scheme would be to classify the texts individually and compare the labellings to determine if the review label requires changing.

The responses submitted to the review by the healthcare providers can be regarded as professional responses, which may feature bias in favour of the health service provider, and therefore not act as reliable indicators of review sentiment. For example, playing down the severity of a complaint or potentially not responding to it at all in the reply. In these cases, the context offered by the response may be minimal. These should not be used as contextual documents in classification, and the relabelling process should be sensitive to this where possible.

If we were to take a probabilistic approach, so only taking the response polarity into account when the review labelling had a low confidence associated with it, then do we could use the response label as the review label. In such an instance, only if the confidence is low would the outcome be recalibrated. The review labelling confidence would be low if the words in the document had not been seen before, or a document of mixed sentiment had been given as input. So the contextual labelling in such a case would provide a stabilising factor.

Similarly, labels may only be commuted if the response confidence is above a given threshold and the review classification labelling below a given threshold. There is a requirement that the classifier is confident in the labelling of the response for it to be of any use. Chapter 6 will discuss the process of recalibration using the review and response labellings in further detail.

Summary

This chapter has discussed the consequences of context on sentiment classification. Rules for opinion identification were shown to be insufficient where the prior polarity of a sentiment expression changed under a given context. This is likened to the challenge of word-sense disambiguation, and the applicability of these to disambiguating word or sentence sentiment is examined. As sentiment classification operates on the document level, we pose the question of whether these methods can be expanded to the disambiguation of document-level sentiment labelling. Previous work on modelling constraints and relationships for sentiment classification are examined, and important works highlighted. In the setting of classifying patient feedback, we find that using the response to a document may prove to be a useful feature in the review's classification. In the next chapter, we examine this intuition further.

CHAPTER 6

SENTIMENT CLASSIFICATION VIA A RESPONSE RECALIBRATION FRAMEWORK

Introduction

A probabilistic classification model outputs the likelihood of a document belonging to a given class. In the binary sentiment classification task, a likelihood greater than 0.5 for a category will assign this polarity as the document-level classification. However, a likelihood of 0.51 is potentially no more useful than a non-biased coin-toss in assigning a polarity to a document and may, therefore, lead to errors in classification. A class confidence greater than 0.5, but substantially lower than 1.0 could occur in the instance that document features convey implicit or ambiguous sentiment. Therefore, a method to recalibrate the classifier outcome would be advantageous if it were to correctly classify the sentiment of a document.

In this chapter, we examine approaches to the recalibration of probabilistic classification models in the presence of context-bearing documents. Three recalibration protocols are proposed and a framework is developed to investigate the application of these conventions on the NCSD. Experimental results reveal significant improvements in classification accuracy over baseline classification methods.

6.1 Motivation

Probabilistic classification models are among the standard supervised machine learning approaches that have been applied to the task of sentiment classification of text (Pang et al., 2002). Given an unlabelled document, a trained probabilistic model is able to determine an appropriate labelling in relation to a given confidence for the proposed labelling. In the two-class sentiment classification problem, a labelling confidence that is greater than 0.5 will lead to a particular sentiment being attached to the input document. However, it is questionable whether a classifier confidence output of 0.51, for example, is sufficiently suitable for the application of any given label. Such a low confidence poses a problem for sentiment classification that could lead to documents being labelled incorrectly to the detriment of a sentiment analysis system. Even some of the higher likelihoods that are still less than 1.0 are indicative of doubt in labelling

A low classifier confidence in sentiment analysis may be produced due to inherent linguistic difficulties that plague systems developed for natural language processing. For example, documents where a sentiment is conveyed implicitly, ambiguously, or in a sarcastic manner can cause problems for machine learning approaches to sentiment classification. Methods have been proposed to deal with such facets of language in sentiment analysis. These tend to focus on hand-crafted lexicons (Balahur et al., 2011) or intra-document contextual cues (Greene & Resnik, 2009) to disambiguate the polarity of a document. We propose a method that takes into consideration related documents in the classification process and duly adjusts classification output using a *sentiment recalibration framework*.

Typical data used to test sentiment classification algorithms consists of a set of isolated documents that convey a sentiment. Amongst others, online reviews are a popular dataset for sentiment analysis algorithm evaluation (Mukherjee & Liu, 2012). Usually, a sentiment classification system will take a document set as input and given an algorithm will classify the polarity of each text. Many different algorithms have been proposed for this purpose, but the general evaluation procedure remains the same. The NCSD is related to the online review domain, but alongside a reviewer's comment, the NCSD contains a document that relates to the review comment, the management response. These acknowledge and respond to the content of the review,

and take into account the sentiment of the comment in the wording of the response. This is of particular use in the medical domain, whereannas Denecke & Deng (2015) note, when people are discussing issues in the clinical domain, language use tends to be more implicit and conservative in order to avoid a misunderstanding that could lead to rash actions or judgements being made. A mechanism that could determine the context of what is being said through an external knowledge source, such as a response, may be useful for this purpose.

Hunston (2011) argues that the concept of evaluation is an "ideology that is shared by writer and reader" and "takes place within a social and ideological framework." The reader and writer may not necessarily agree on a sentiment-bearing proposition, but the reader's interpretation of the writer's opinion is an important notion that underpins sentiment analysis. Inter-annotator agreement studies (Wilson & Wiebe, 2005; Wilson, 2008a; de Kauter et al., 2015) have been undertaken in order to demonstrate some semblance of agreement with notions of sentiment and its conveyance in text; a judgement on what the annotators mutually believe the writer's true sentiment was in light of their review. If this annotated data is used as training data, systems are then developed to spot identifiers of sentiment based on sentiment expressions in text that annotators have mutually agreed on. In such a system, common knowledge introduced through annotation is an indicator of a shared value system. This can be taken a step further. Instead of just allowing a discrete, potentially binary annotation for labelling sentiment-bearing documents, the annotator could instead be given the ability to give more detail on their thoughts that motivate their annotation and allow them to give a detailed response to the initial comment. This response should mimic the rating of the initial comment, but also, be more explicit about the sentiment of the original comment. Now, this process could go on ad infinitum, with annotators annotating annotations. In this study, we choose to cap the level of annotation to just one for the present study, but this is something that could be explored further.

Our proposed method takes into account external but relevant documents during the sentiment recalibration process. We use these documents to make adjustments to classifier outputs, in an adjustment and correction phase. To our knowledge, this is the first work in sentiment classification to attempt the recalibration of a sentiment classifier given relevant documents.

We attribute this ability to the format and depth of the dataset used in this thesis, the NCSD.

Currently examined approaches to recalibration may typically rely on Platt scaling or binning methods. Platt scaling trains a logistic regression model on the output of an SVM classifier, enabling the production of posterior classification probabilities (Platt, 1999). Binning is another calibration method that is particularly effective for classification (Zadrozny & Elkan, 2001). Such recalibration methods focus on statistical methods of recalibrating classifier output. However, when dealing with related natural language documents, we can use inferences from the content of the related text to guide the recalibration process. Therefore, we propose the use of the response to recalibrate the labelling of the initial comment. This takes a response directed at a comment and uses the outcome of its classification as a starting point for recalibration. We discuss in further detail the recalibration protocols in the following sections.

Work on bagging in sentiment classification is somewhat related to our work (Dai et al., 2011; Nguyen et al., 2013). Bagging trains a number of models on a similar set of training data. During classification, each model then classifies the given instance, and a voting protocol labels the instance with the majority label suggested. However, our framework does not train multiple classifiers, although there is the potential for the framework to be extended to incorporate this. Instead, a related document is used to guide and recalibrate the outcome of the initial classification. Our method does not suffer from the issue of low classifier trustworthiness, as we will demonstrate in the response classification baselines. The need for further methods; such as stacking (Wolpert, 1992), which is a method that combines multiple classification models and is applied where low classifier trustworthiness is an issue, can therefore be eliminated.

6.2 Research Question

The following research question will be examined in this chapter:

RQ5: To what extent can a review response be used to improve the classification results of a collection of patient feedback?

In the previous chapter, the argument that a response attaches a context to a review was presented. In Chapter 3, the vocabulary of the responses was shown to be more concise than that of the reviews. Additionally, in this chapter, high frequency and key terms were found to be indicative not of the response sentiment, but the sentiment of the review. The constrained nature of the responses may indicate that in the clinical domain these may be suitable for recalibrating review labelling given low likelihood labelling.

This question will be answered by developing an experimental framework to test the hypothesis that review responses can be used to recalibrate the outcome of a classifier trained to categorise a document set of reviews by their sentiment. Results that improve upon a baseline would indicate that improvements can be yielded. Multiple classifiers will be tested upon the best performing feature set from Chapter 4.

In order to approach this research question, In this work, four assumptions are made regarding the initial review comment and its relative response:

1. The responder has read the review.
2. The responder responds to both the content and sentiment of the review.
3. The manner that a responder replies to a negative review differs from the way that they reply to a positive review.
4. The difference in response can be used to identify the sentiment of the review.

By making these assumptions, an experimental framework can be developed to evaluate the recalibration potential that responses have for the task of sentiment analysis.

Algorithm 1 Recalibration framework

global variables

$mlAlgo$, global var1

end global variables

function FRAMEWORK(D)

$C, R \leftarrow$ EXTRACTSUBDOCS(D)

$comTrainingData, comTestData \leftarrow$ SPLIT(C)

$respTrainingData, respTestData \leftarrow$ SPLIT(R)

$mlAlgo \leftarrow$ SELECTMLALGORITHM

$comModel \leftarrow$ TRAINCLASSIFIER($comTrainingData$)

$respModel \leftarrow$ TRAINCLASSIFIER($respTrainingData$)

RECALIBRATE($comModel, comTestData, respModel, respTestData$)

6.3 Methodology

To examine the research question, an experimental framework was developed to evaluate the recalibration protocols. The framework is described at a high level in Algorithm 1, and in this section, the experimental framework will be detailed further.

The FRAMEWORK first requires a document set D as input. In the case of our experiments, input is the NCSD. From this, the two main document sets for the experiments are extracted: the review comments C and their responses R. These are then further split into training sets consisting of 75% of the input data, $comTrainingData$ and $respTrainingData$, and the remaining documents form the test sets, $comTestData$ and $respTestData$. Following this, the machine learning algorithm for that particular iteration of the evaluation is chosen. A global variable $mlAlgo$ is set by the method SELECTMLALGO to one of the three supervised machine learning algorithms selected for experimentation: Naive Bayes (NB), Multinomial NB (MNB) or support vector machine with Platt scaling to produce a probabilistic output (SVM). These are chosen on the basis of the results of Chapter 4. While classification models exhibited little

statistical difference in overall performance, examination of the rankings shows that the MNB method consistently performed highly in previous experiments, and the SVM and NB models ranked as two of the poorer learning algorithms when classifying patient feedback by the sentiment conveyed. The latter two models were chosen in order to determine the potential for recalibration to improve the performance of these classification models to compete with the top ranking MNB classifier. Classification models are invoked using the Weka Toolkit (Hall et al., 2009). The chosen classifier is stored in the global variable as it remains constant throughout that particular run of the experiment, where it is used in both the TRAINCLASSIFIER method in the framework, and the CLASSIFY function of the recalibrate methods that will be discussed later in this section. Two classification models, $comModel$ and $respModel$ are then trained using the selected classification method and the respective training data. These models, and the respective test data sets are then passed to the RECALIBRATE method to undergo appropriate recalibration through the application of the proposed recalibration protocols.

6.3.1 Protocol overview

Incorporated into the classification framework is the RECALIBRATE function. The following three recalibration protocols are proposed for examination through this method:

1. **Probabilistic threshold recalibration**: we experiment with the transference of response labels to their corresponding reviews at varying comment labelling confidence levels as defined by the experimental framework threshold. We transfer the label if the comment labelling confidence is at or below the experimental framework threshold.

2. **Strong probabilistic threshold recalibration**: this follows the above protocol of transferring the response label to the comment. However, in this version an additional constraint is imposed, and the label is transferred from the response to the review only if the labelling confidence of the classified response is greater than or equal to the confidence of the labelling of the review.

3. **Document similarity recalibration**: the label is transferred from the response to the

review if a level of document similarity is exhibited between the response and review. Classification confidence levels are not considered when applying this protocol due in order to study the effects of document similarity independently of the effect of classifier confidence.

These are detailed further in the following subsections.

6.3.2 Probabilistic threshold recalibration protocols

Algorithm 2 Probabilistic threshold recalibration protocol

function RECALIBRATE($comModel, comTestData, respModel, respTestData$)

$threshold \leftarrow 0.5$

while $threshold \leq 1.0$ **do**

for all $c, r \in comTestData, respTestData$ **do**

CONF(c) $\leftarrow$ CLASSIFY($comModel, c$)

CONF(r) $\leftarrow$ CLASSIFY($respModel, r$)

if CONF(c) $\leq threshold$ **then**

SENT(c) $\leftarrow$ SENT(r)

EVALUATERECALIBRATION($comTestData$)

$threshold \leftarrow threshold + 0.01$

Classifier confidence is an important aspect of sentiment analysis. The probability of a label being assigned to a document with a certain category is just as important as the labelling itself. A label may be assigned, but a weak confidence could be associated with the labelling decision. In this scenario, an external but relevant source of information regarding the sentiment of the document could be used to recalibrate the outcome of the initial classification where review labelling was not fully confident.

Algorithm 2 formalises this notion of classifier recalibration given the possibility of low classification confidence. The RECALIBRATE method for the probabilistic threshold recalibration protocol is therefore developed to investigate whether the response labelling can be used to

improve the performance of the standard machine learning models applied to the task of sentiment classification in the clinical domain. It is not known at what confidence level classifier recalibration would be effective, hence in this protocol, the method steps through each potential confidence value, starting at the minimum confidence value of 0.50 for a particular review labelling, and iterating through steps of 0.01 to a maximum confidence value of 1.0. For each step, referred to as $threshold$ in algorithm 2, all test data is classified using the CLASSIFY method by applying the classification model stored in the global variable $mlAlgo$ of the framework. If it is found that the confidence of the labelling of the review comment, CONF(c), is less than or equal to the threshold value, then the sentiment label of the response, SENT(r), is used to replace the initial sentiment labelling of the comment, SENT(c). The performance of the recalibration is then evaluated using the EVALUATERECALIBRATION method in the last call of each iteration of the while loop.

This process changes when implementing the strong probabilistic threshold recalibration protocol; detailed in Algorithm 3. In this, not only is there a requirement that CONF(c) is less than or equal to the $threshold$ value of the current iteration, but also that the confidence of the response labelling, CONF(r) is also greater than or equal to CONF(c).

Both protocols rely on the outcome of the response classification being reliable. The agreement study in Chapter 3 between comment sentiment and response sentiment yields $\kappa = 0.761$, a good level of agreement. This agreement is not any higher as a number of positive comments are replied to by responses that appear to acknowledge an aspect of negative sentiment. However, this begs the question of whether a stronger probability is required to recalibrate the comment sentiment, hence the formation of the strong probabilistic threshold recalibration protocol.

Algorithm 3 Strong probabilistic threshold recalibration protocol

function RECALIBRATE($comModel, comTestData, respModel, respTestData$)

$threshold \leftarrow 0.5$

while $threshold \leq 1.0$ **do**

for all $c, r \in comTestData, respTestData$ **do**

CONF(c) $\leftarrow$ CLASSIFY($comModel, c$)

CONF(r) $\leftarrow$ CLASSIFY($respModel, r$)

if CONF(r) $\geq$ CONF(c) **AND** CONF(c) $\leq threshold$ **then**

SENT(c) $\leftarrow$ SENT(r)

EVALUATERECALIBRATION($comTestData$)

$threshold \leftarrow threshold + 0.01$

6.3.3 Document similarity recalibration protocol

One of the assumptions made when applying the previous probabilistic thresholding recalibration protocol is that the response is always relevant to the review that it is replying to. However, this notion of relevance between review and response is not guaranteed. In the exchange, as with most natural dialogue, no constraints exist stipulating that the reply must respond to any of the points introduced in the review. Responses of this nature may just have been given as a formality, and due to this, the content of the response may not be indicative of the responder having acknowledged the feedback given in the review. The nature of the lack of relevance in a responder's utterance has been discussed in some detail by Grice (1970) and Ginzburg (2010), and they propose that a lack of relevance may be indicative of an unwillingness to address the prior utterance.

In the recalibration framework, we examine the use of the response as a recalibrating factor for the review classification. A generic response such as the one discussed above is not indicative of comment sentiment, and would be of little use as a recalibrating factor. However, figure 6.1 shows a review-response pair, whereby the review sentiment is ambiguous. In this example, the response clarifies the sentiment of the review in its wording. It makes the sentiment clear

Review Sentiment:Ambiguous

The nurses really gave it their all.

Response Sentiment:Positive

Thank you for your kind comments about the nurses, we are pleased that you feel this way. They always give it their all!

Figure 6.1: Example review and response whereby the response is indicative of the review's sentiment.

through the phrases *kind comments* and *we are pleased*. Importantly, there is a lexical overlap between the review and the response on the phrases *the nurses* and *it their all* that indicates the potential for a shared sentiment. Therefore, we investigate the possibility that if the response shows signs of similarity to the review, then it should be used for recalibrating a classification choice. A computational concept of relevance between a response and review has not before been explored so it is unknown what level of similarity is required for appropriate recalibration. The thresholding framework proposed for the previous probabilistic protocol is therefore of use in investigating this concept.

In search, relevance is characterised as the similarity between a query and the documents of the search space (Manning et al., 2008). The greater the overlap between the query and the documents, the higher the likelihood that the documents are topically relevant to the query. This intuition frames relevance as the ability to compute the lexical similarity between texts. Lexical similarity is often studied from the perspective of an information retrieval (Strzalkowski, 1995), plagiarism detection (Chong & Specia, 2012) or authorship attribution task (Luyckx & Daelemans, 2008) in natural language processing, and not the perspective of inter-document correspondence. We make the assumption that if the review and response exhibit traits of lexical similarity, then the response, as it is generated following the submission of the review, is relevant. If it is relevant, then we can also assume that it is responding to the sentiment of the review, as this is inherent to the nature of the review response.

The cosine similarity (Mihalcea et al., 2006) and Greedy String Tiling (GST) (Wise, 1993) algorithms have been used to compare document similarities for the measurement of text reuse (Clough et al., 2002) and semantic similarity between documents (Pilehvar et al., 2013). Both approaches are able to output a value denoting the level of surface similarity between documents. By examining the similarity between a review and a response, we are not explicitly looking to retrieve or determine the source of a text, but only determine that the response addresses the review. The GST algorithm is flexible to the transposition of strings in a document but limited enough so that the order of words in a document is not lost when determining similarity, unlike the cosine similarity approach.

Given two documents, A and B, the GST algorithm computes the longest matching substrings between them. It does this through a process of tiling, whereby a tile is a unique match of a substring from A with a substring from B. Through a process of iteratively scanning and comparing the words of each document to find maximal matching patterns, once a tile is found, it is marked and cannot be used in further comparisons. Importantly, a minimum-matching length can be declared when initialising the GST algorithm to indicate that tiles must be of at least a certain length. Any tiles below this set length are then ignored and do not contribute to the overall similarity calculation. For example, setting the minimum-matching length to two will mean that all tiles of length one, single words, will be ignored if they are found to be maximal matching substrings. The GST algorithm is defined in algorithm 4.

The GST algorithm begins with an empty list, $matches$, that will eventually hold any matching tiles that are found. The variable $lengthOfTokensTiled$ is also initially set to zero. During the body of the main loop, the variable $maxMatch$ is assigned the value of the $minimumMatchLength$. This could change of over the course of the algorithm, but the main loop terminates if $maxMatch$ is found to still match the $minimumMatchLength$ at the end of the main loop's body. Two inner for loops then iterate over the tokens of the first document P, and the second document T, if these are unmarked. Tokens are marked when they are found to appear in a tile, and these cannot be used for further matches. Within these for-loops, if a token P_p is found to match a token T_t and both are unmarked, the inner while-loop attempts to match

Algorithm 4 Greedy string tiling algorithm

```
function GREEDYSTRINGTILING(P, T, minimumMatchLength)
    matches ← {}
    lengthOfTokensTiled ← 0
    repeat
        maxmatch ← minimumMatchLength
        for all UNMARKED(P_p) ∈ P do
            for all UNMARKED(T_t) ∈ T do
                j ← 0
                while P_{p+j} == T_{t+j} AND UNMARKED(P_{p+j}) AND UNMARKED(T_{t+j}) do
                    j ← j + 1
                if j == maxMatch then
                    matches ← matches ⊕ MATCH(p, t, j)
                else if j > maxMatch then
                    matches ← {MATCH(p, t, j)}
        for all MATCH(p, t, maxMatch) ∈ matches do
            if NOTOCCLUDED then
                for j ← 0, maxMatch − 1 do
                    MARKTOKEN(P_{p+j})
                    MARKTOKEN(T_{t+j})
                lengthOfTokensTiled ← lengthOfTokensTiled + maxMatch
    until maxMatch == minimumMatchLength
```

Figure 6.2: Greedy String Tiling Algorithm (Wise, 1993)

the longest possible token sequence until it is no longer possible to match any further, due to a difference in tokens occurring, or a marked token being discovered in the sequence. Following this, if the length of the sequence is found to be equal to the current value of $maxMatch$, then that match forms a triple MATCHES(p, t, j), where p and t are the starting indices of the matching tokens in their respective documents, and j is the length of the match. However, if j is found to be greater than the current value of $maxMatch$, $matches$ is re-initialised as a new list holding only the current match of maximally discovered length. This process should find the longest match for that run of the for-loops.

Following the completion of the first set of for-loops to determine the maximal-matches for the given pass of the algorithm, a second for-loop is used to investigate whether any of matching token strings held in $matches$ occludes a token stored in a tile that was created earlier. If it is the

case that the observed match triple doesn't, then all tokens in that match are marked. Following this process, the variable $lengthOfTokensTiled$ is increased by the value of $maxMatch$ for each non-occluding match.

This process continues until $maxMatch$ equals the predefined $minimumMatchLength$, when the program terminates.

On completion of the GST algorithm, a final length of the list of tokens tiled can be accessed. This can be used to calculate a numerical level of similarity between A and B. Similarity can be calculated on the basis that a similarity value of 0 indicates that the two document strings were not equivalent, and a similarity of 1 indicates that the document strings were equivalent. Following this heuristic, Prechelt et al. (2000) calculates a similarity value that takes into account the coverage of the tiles in respect to the length of the documents that are being compared for similarity:

$$sim(A, B) = \frac{2 \cdot coverage(tiles)}{|A| + |B|} \tag{6.1}$$

$$coverage(tiles) = \sum_{match(a,b,length) \in tiles} length \tag{6.2}$$

where $sim(A, B)$ is the similarity score between A and B, $coverage(tiles)$ is a numerical value denoting the sum of the length of all the tiles that are found, and $match(a, b, length)$ is a representation of a tile, and is a triple denoting an association between identical substrings of A and B, starting at position a in A and b in B, where both matching substrings are of the same $length$.

Clough (2003, p.138) suggests calculating a containment score in respect of the coverage of the tiles in comparison to the length of document B:

$$containment(A, B) = \frac{2 \cdot coverage(tiles)}{|B|} \tag{6.3}$$

Again, this yields a value between 0 and 1, but in this case, instead of being normalised by the sum of both lengths, this metric would calculate the proportion of tiles detected in the

Similarity:0.058

Review

if the practice wishes to continue to offer an online repeat prescription request service, i suggest they do something to ensure that the system works.

Response

thank you for adding a comment about our practice. we are pleased that you rate the care provided by our medical staff as excellent. we are sorry that you encountered a problem with the online repeat prescription service. it would help us if you could contact our manager joy hancock directly so that we can investigate what went wrong in this instance. we are currently overhauling our website from scratch. this should make it much more user-friendly and should reduce the possibility of any such errors happening. we are also in the middle of a major refurbishment and extension project at the surgery. this will result in more space for patients and staff, extra consulting rooms and improved disabled access and facilities. there will also be a pharmacy on the ground floor of the extension. we hope that this will improve the overall experience for patients at our surgery. thanks again for your comment.

Figure 6.3: An example tiling with a low similarity between the response and the review.

response, and hence for the work in this thesis, gives a more appropriate value that denotes the relevance of the response to the review based upon the level of document similarity.

Despite the similarity score being capable of falling between 0 and 1, it is unlikely that a response would exactly mirror the language of a response, and yield a 'perfect' similarity score of 1.0. An initial pass over the data using the GST with a minimum-matching length of 1 yields a range of scores from 0 to 0.5. Due to this, we define low similarity as a score less than 0.1, a mid-level similarity between 0.1 and 0.3, and a score denoting a high level of similarity sitting between 0.3 and 0.5. Examples of review-response pairings that exhibit these different similarity levels are given in figures 6.3 to 6.5.

The order preserving nature and flexibility of the GST algorithm are desirable traits when calculating relevance between response and review, and due to this we implement a simple version in the document similarity protocol that calculates the similarity score that normalises

Similarity:0.154

Review

going [to] [the dentist is] never a [pleasant] experience but [i would] [have] no hesitation in recommending this dentist, they are without doubt [the] best i have visited over [the] years.

Response

dear patient, on behalf of [the] practice team [i would] like [to] thank you so much for your lovely comments that you [have] made about our practice. it was really nice to see that although attending [the dentist is] not [the] most [pleasant] of experiences, we were able to make your visit as comftable as possible. we look forwarding to seeing you again soon.

Figure 6.4: An example tiling with a mid-level similarity score between the response and the review

the length of the tiles by the length of the response document.

Algorithm 5 outlines the document similarity protocol and follows a similar logical flow to the algorithms proposed for the probabilistic threshold recalibration methods. However, instead of the $threshold$ being instantiated at 0.50, in this protocol it is instead instantiated at 0.0. The same 0.01 increments still apply, and the function iterates until a value of 1.0, for completeness. The central loop iterates over all of the test documents, and the method CALCULATEDOCUMENTSIMILARITY takes a comment and its adjoining response and calculates the level of similarity using a call to the GST method described previously that given the length of the tiles is able to produce a similarity score. SIMSCORE is used to represent the resulting similarity score between c and r. If this is found to be greater than or equal to the current threshold value of that particular iteration then SENT(c) is replaced by the labelling SENT(r). At the end of each iteration of the outer loop, the method EVALUATERECALIBRATION monitors the performance of the recalibration protocol at the given threshold value. Results are recorded and presented graphically in the following section.

Similarity:0.402

Review

visited for my pill check and after waiting an hour i was seen by the nurse in the waiting area as they had someone booked in who needed a lengthy consultation. telephoned today on the advice on nhs direct for an urgent appointment for my 2 year old son who has recently been in hospital and had been unwell for nearly a week (having already been seen by the hospital and walk in centre as we can never get an appointment) to be told they were fully booked and could do nothing. the staff are unhelpful and i have been told that they are required to keep back a certain number of appointments every day for emergencies and would have thought a 2 year old with recent health issues and on the advice on a medical worker would have fallen under this category.

Response

we are very sorry for your experience on trying to get an appointment for your 2 year old son .we apologise that you were kept waiting for your appointment.we do aim as a surgery to provide the best level of care and services to our patients. we do welcome feeedback both positive and negative. we do in the majority of occassions see children under the age of 4 years on the day . please do make an appointment with the surgery senior gp manager to discuss this further should you wish.

Figure 6.5: An example tiling with high similarity score between the response and the review.

Algorithm 5 Similarity threshold recalibration protocol

function RECALIBRATE($comModel, comTestData, respModel, respTestData$)

$threshold \leftarrow 0.0$

while $threshold \leq 1.0$ **do**

for all $c, r \in comTestData, respTestData$ **do**

SIMSCORE$(c, r) \leftarrow$ CALCULATEDOCUMENTSIMILARITY(c, r)

if SIMSCORE$(c, r) \geq threshold$ **then**

SENT$(c) \leftarrow$ SENT(r)

EVALUATERECALIBRATION($comTestData$)

$threshold \leftarrow threshold + 0.01$

6.4 Evaluation

Each of the experiments will be compared to a baseline method that consists of the particular classification model under consideration with no recalibration applied to it. Results from the recalibration experiments will be tested for statistical significance using the McNemar test (McNemar, 1947) that examines the marginal homogeneity of the classifiers for equality; that is, whether the recalibration protocols are able to produce fewer errors when examining the differences between the baseline and the recalibrated methods. We adapt the description given by Dietterich (1998) of the McNemar test for machine learning to refer to the baseline classifier and a given a classifier that has undergone recalibration. In his description of the test, the following four values are first observed:

1. The number of documents misclassified by both classifiers (n_{00})

2. The number of documents misclassified by the baseline classifier but correctly classified during recalibration (n_{01})

3. The number of documents misclassified during recalibration but correctly classified by the baseline classifier (n_{10})

4. The number of documents correctly classified by both the baseline classifier and the recalibrated classification outcome. (n_{11})

Given these values, the McNemar statistic is then calculated as follows:

$$\frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}} \tag{6.4}$$

The value yielded from the test is compared to the χ^2 distribution table with one degree of freedom, where $p < 0.01$. If the calculated McNemar statistic is greater than 6.635, the null hypothesis that the recalibrated method and the baseline classifier perform comparably can be rejected. Baseline results for comparison in the evaluation of the experiments are shown in table 6.3, and in figures 6.6 to 6.25, these are represented by the olive, horizontal lines for comparison.

6.4.1 Response classification

To demonstrate the relative simplicity in which responses can be classified to an adequate degree, a simple rule-based classification system using a lexicon of fourteen words that identify the sentiment the response is replying to. The lexicon was generated through a manual selection process following a keyword analysis procedure, detailed in Chapter 3. Reducing some of these words to their stems, shown in Table 6.1, a recall value of 0.900 is achieved. The response sentiment label is compared to the gold-standard labelling of the advice, producing an accuracy of 63.341%.

Following this, the reliability of response classification is examined using supervised classification models. This incorporates a ten-fold stratified cross-validation in the Weka toolkit to determine the results of response classification using the NB, MNB and SVM supervised learning models. Results are reassuring, and exceed those achieved in Chapter 4 for review classification. We discussed the traits of the responses in Chapter 3, but to reiterate, the relatively focused vocabulary of the responses given the number of documents contributes to this result. The responses do not tend to deviate from their given scripts and therefore supervised

Table 6.1: Positive and negative word stems used to identify sentiment-sensitive responses

Positive	Negative
apprecia*	apolog*
deligh*	sorr*
posit*	sad*
good*	improv*
kind*	unhapp*
pleased	serious*
	concern
	discuss

Table 6.2: Response baseline classification results (+1 = positive -1 = negative)

	Accuracy	Precision	Recall	F_1
NB +1	0.906	0.749	0.909	0.821
NB -1	0.906	0.97	0.905	0.937
MNB +1	0.944	0.841	0.94	0.888
MNB -1	0.944	0.981	0.945	0.963
SVM +1	0.951	0.884	0.913	0.899
SVM -1	0.951	0.973	0.963	0.968

learning models fair well when the training and test data are similar. Table 6.2 shows the results of response classification using the aforementioned supervised machine learning models. The best accuracy is 0.951, achieved using the SVM classification model. These results show the reliability of classifying the responses and the high degree of accuracy when using these labels for review classification recalibration.

6.4.2 Results: Probabilistic threshold recalibration

Figures 6.6 to 6.10 report the results of the first recalibration protocol, probabilistic threshold recalibration.

Table 6.3: Type 2 review baseline classification results (+1 = positive -1 = negative)

	Accuracy	Precision	Recall	F_1
NB +1	0.676	0.883	0.645	0.7454
NB -1	0.676	0.436	0.761	0.554
MNB +1	0.860	0.911	0.898	0.904
MNB -1	0.860	0.727	0.756	0.741
SVM +1	0.816	0.901	0.841	0.870
SVM -1	0.816	0.628	0.746	0.682

Results highlight an improvement in classifier accuracy as the probability threshold increases for all learning models tested. All start the recalibration process at the threshold value, but gains can be found at various threshold value beyond a probability threshold of 0.51. The MNB instantly improves on classification accuracy, and it remains above the baseline for the remainder of the iterations, reaching a peak at 0.88 of 91.408%, which corresponds to an F_1 of 0.902. The SVM does not improve upon the baseline until a confidence of 0.58, and peaks at 0.98, with an accuracy of 86.300%, before results decrease at greater probability thresholds. The NB crosses the baseline at a confidence of 0.51 and rapidly increases in accuracy from 0.713 at a confidence of 0.99 to 0.856 at a full relabelling when all labels are used for recalibration.

The precision results for the negative class shows similar traits to the accuracy results, whereby all models show considerable improvements, highlighted in Figure 6.8, with recalibration pushing both the NB and SVM results beyond the MNB baseline. However, this comes at the cost of the positive precision, where minimal improvements over each classifier's respective baseline are yielded (figure 6.7). The recall results for each classifier shows significant improvements for the positive class for each model over its respective baseline, whereas, for the negative class, recall results cluster around each classifier's respective baseline, with results dropping as the probability threshold reaches 1.0, as shown in figures 6.9 and 6.10.

The number of candidates for recalibration increases as the probability threshold increases,

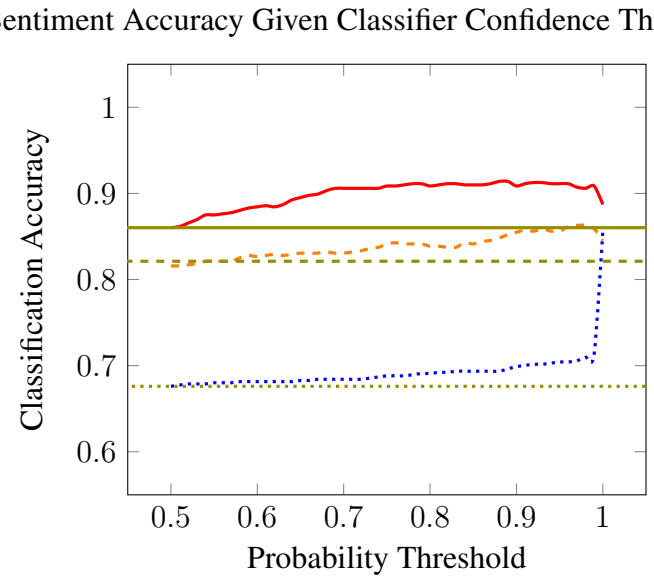


Figure 6.6: Graph of sentiment accuracy results given classifier confidence.

and the number increasing drastically as the probability threshold tends towards 1.0 (Figure 6.11). At the beginning of the recalibration process, the ratio of relabelled instances that correctly relabel the review sentiment is relatively high for each model, shown in figure 6.12, but as more candidates for recalibration are introduced as the probability threshold increases, the success ratio decreases to approximately 0.9 for all examined models.

If we test all models for significance against the baseline results using the McNemar Test, we find that for the MNB classifier between the probability thresholds of 0.54 and 0.99 there is a significant increase in classification performance beyond the baseline values ($p < 0.01$). For the NB classification model, a significant increase in classification performance is not reached until the 0.78 threshold, and significant increases continue as the threshold tends towards 1 ($p < 0.01$). For the SVM model, two zones are found where statistically significant increases in performance beyond the baseline method occur: between 0.73 and 0.81, and 0.84 to 0.99 ($p < 0.01$).

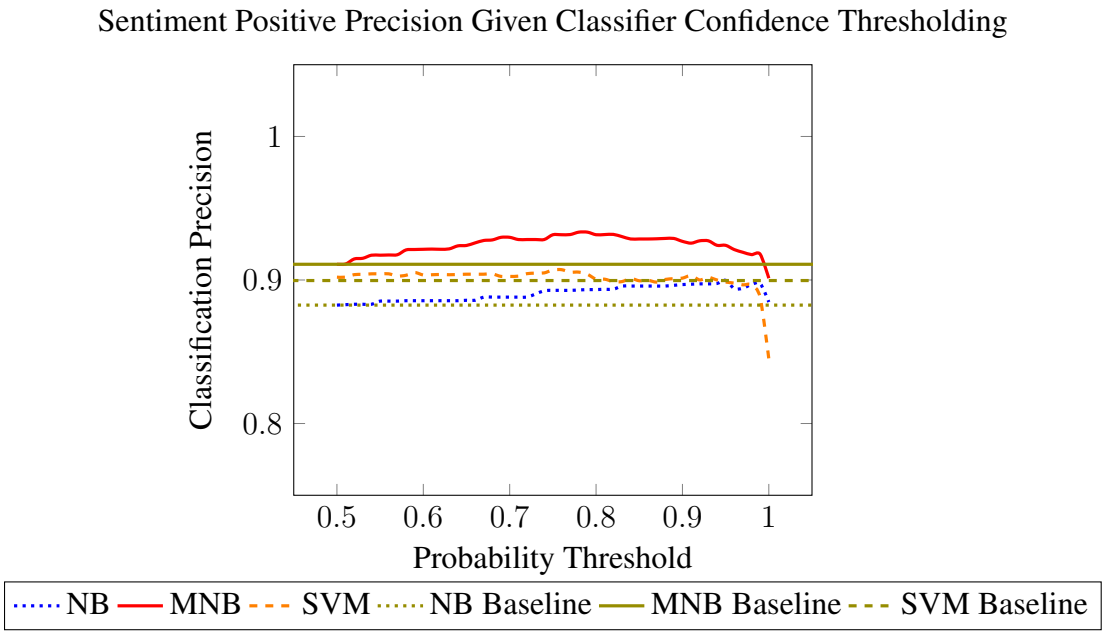


Figure 6.7: Graph of positive class precision results given classifier confidence.

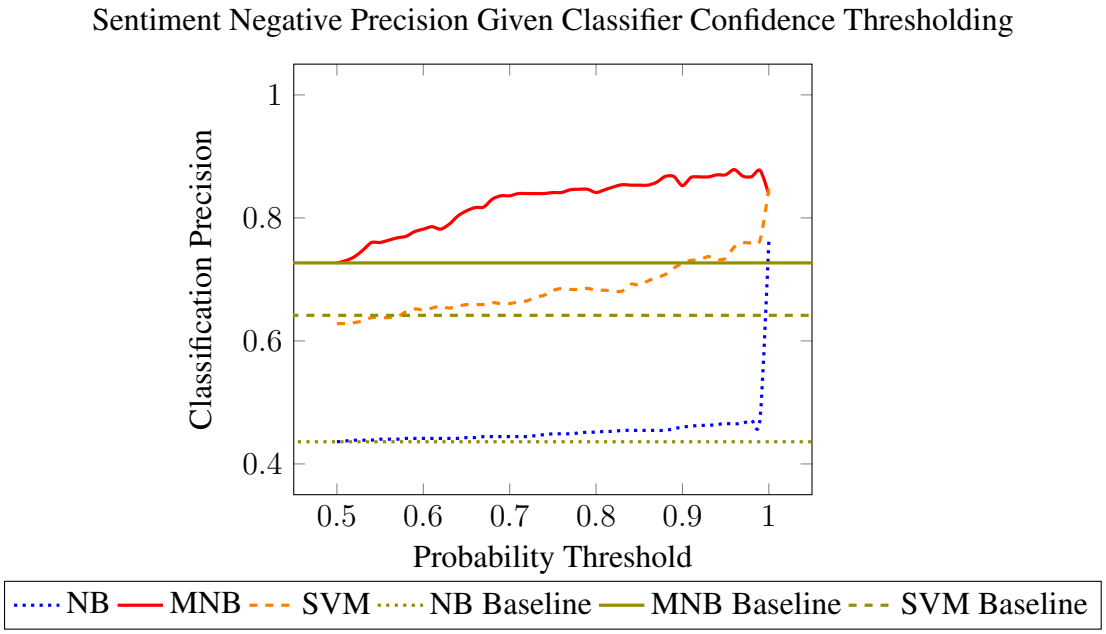


Figure 6.8: Graph of negative class precision results given classifier confidence.

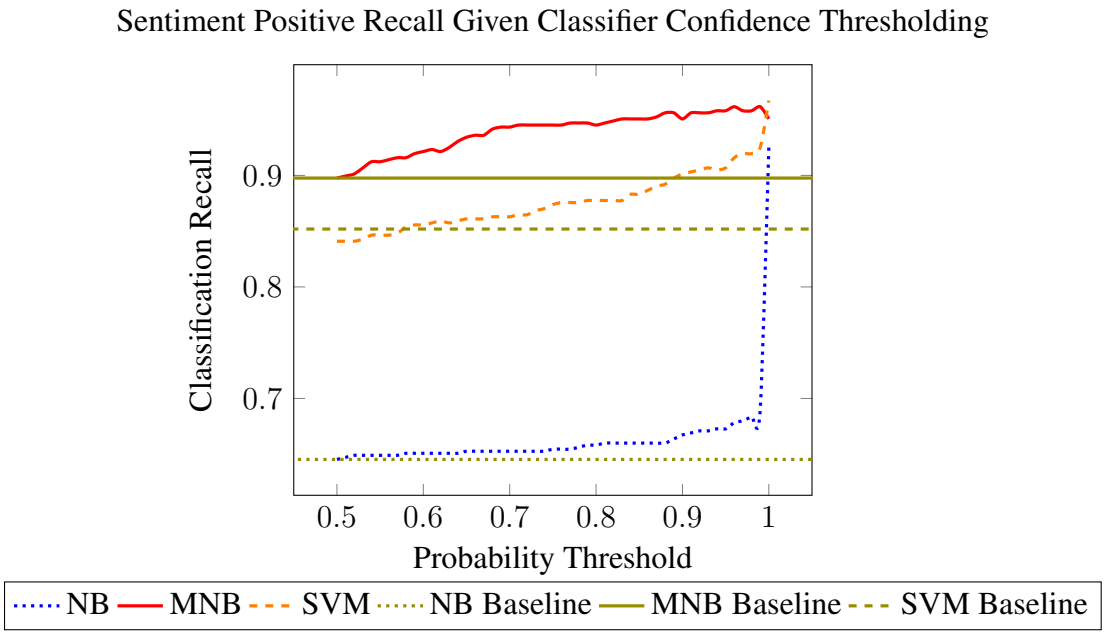


Figure 6.9: Graph of positive class recall results given classifier confidence.

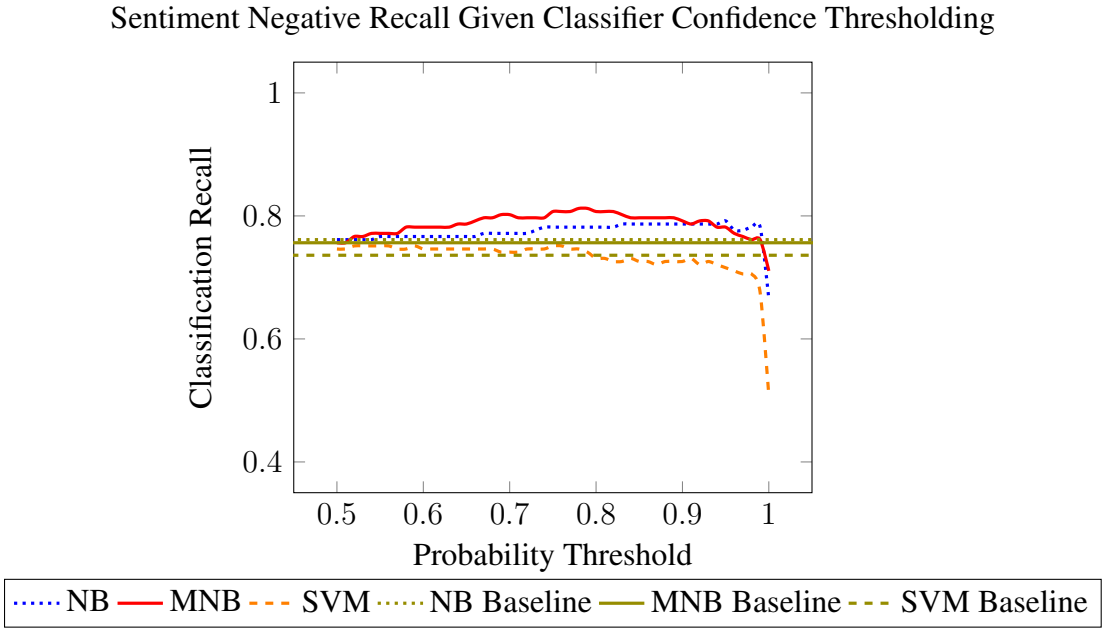


Figure 6.10: Graph of negative class recall results given classifier confidence.

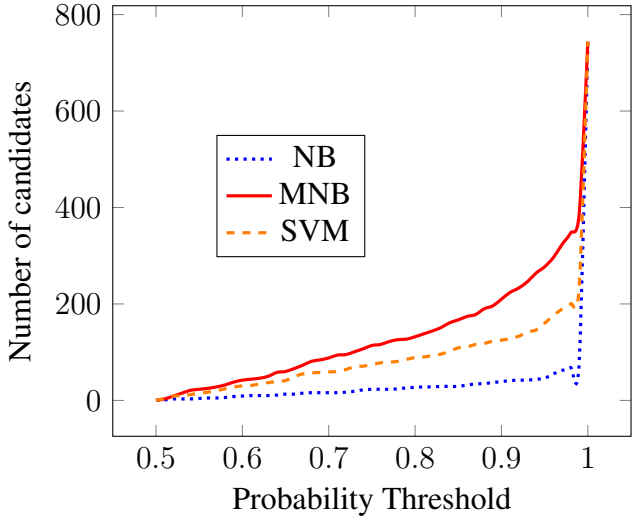


Figure 6.11: Relabelling candidates given varying classifier confidence thresholds.

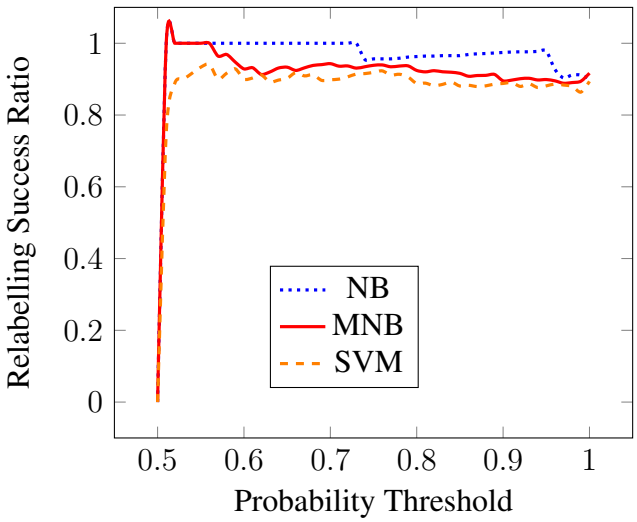


Figure 6.12: Relabelling success rate given varying classifier confidence thresholds.

6.4.3 Results: Strong probabilistic threshold recalibration

The strong response classification experiments impose a constraint that labelling of the comment can only be used for recalibration from the response classification if and only if the response classifier confidence is equal to or higher than that of the comment classification of a given instance. The constraint appears to give a stabilising quality to the outcome of sentiment classification of the reviews.

Comparing figures 6.6 and 6.13, the strong probabilistic threshold yields a substantially smoother gradient to the curve in comparison to the general probabilistic threshold experiments. However, in the strong probabilistic threshold, the steep ascent of the NB classifier appears to have been lost by imposing this constraint, and the accuracy barely surpasses the baseline. The maximum accuracy yielded is found using the MNB model, and is 89.893% at a threshold of 1.0, which does not outperform the 91.408% achieved at a threshold of 0.88 using the same model for the probabilistic threshold recalibration protocol.

Comparing the precision and recall results shown in figures 6.14 to 6.17 to the results from the first protocol, the strong probabilistic threshold recalibration protocol does not tend to yield results that are better than those of the first protocol. As shown in figure 6.16, positive recall does not tail off at higher probabilities for the SVM model in the strong probabilistic threshold recalibration, whereas for positive class recall shown in figure 6.7 for the normal probabilistic threshold calibration, it does.

When comparing the number of candidates available for the strong probabilistic threshold recalibration protocol shown in figure 6.18 to the standard protocol, the NB classifier does not yield as many candidates as the other models. For example, at a probability threshold of 0.99, there are 79 candidates for recalibration, given the constraints of this protocol, compared to 362 for the MNB model, and 234 for the SVM classification model. Moving to a threshold of 1.0, the NB jumps to 597 candidates, the MNB to 637, and the SVM to 733. This indicates the relative confidence of the SVM in its classification, with only 11 potential response labels having a confidence less than the confidence labels of the comments. Despite this, as shown in figure 6.19, the success ratio is still competitive at a high probability threshold despite the influx

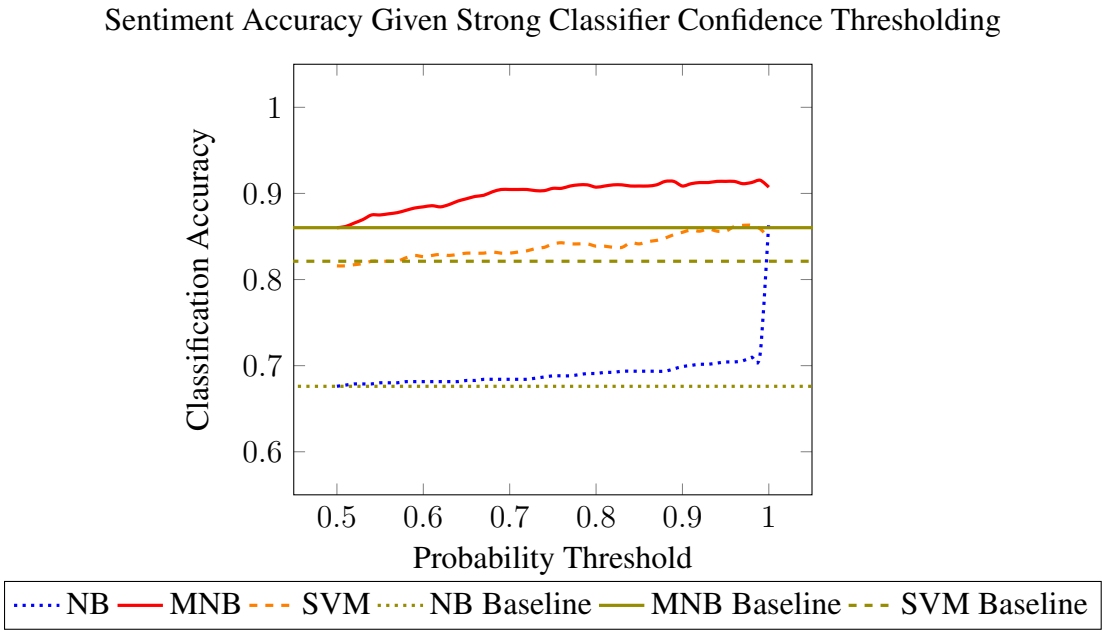


Figure 6.13: Graph of sentiment accuracy results given strong classifier confidence.

of extra potential recalibrating response labels.

Comparing the methods to the baseline classification methods, the results again tend to mirror those of the probabilistic threshold recalibration protocol. However, there is a slight difference in results from the MNB model whereby use of the strong probabilistic recalibration protocol significantly increases the performance over the baseline method where the probability threshold is 1.0, which is not the case in the standard probabilistic threshold recalibration protocol. This is as in the current protocol the accuracy is 1.19% higher than the first protocol. The results from the NB model follow those of the first protocol, reaching a significant improvement at a probability threshold of 0.78 and carrying on until the threshold reaches 1.0 ($p < 0.01$). The SVM models also follow the same trend of the first protocol, where the two zones of significant improvement to the classifier performance over the baseline is achieved between 0.73 and 0.81, and 0.84 to 0.99 ($p < 0.01$).

While a stricter overall protocol than the first protocol, the application of the strong probabilistic threshold does not yield better results than the initially proposed protocol, although statistically significant improvements above the baseline can be seen throughout the aforementioned results.

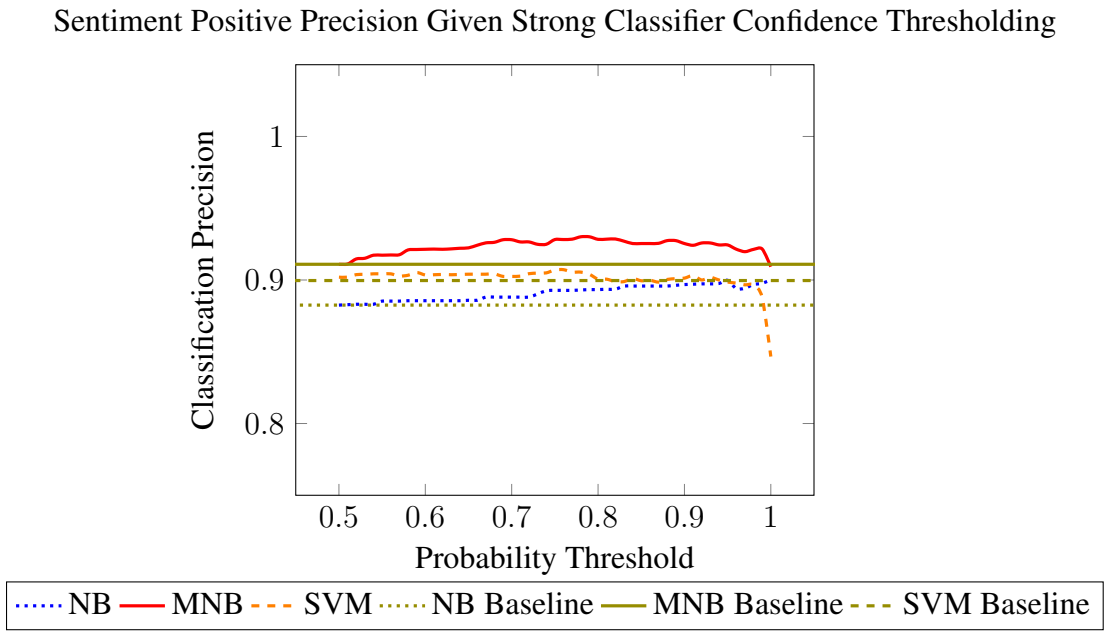


Figure 6.14: Graph of positive class precision results given strong classifier confidence.

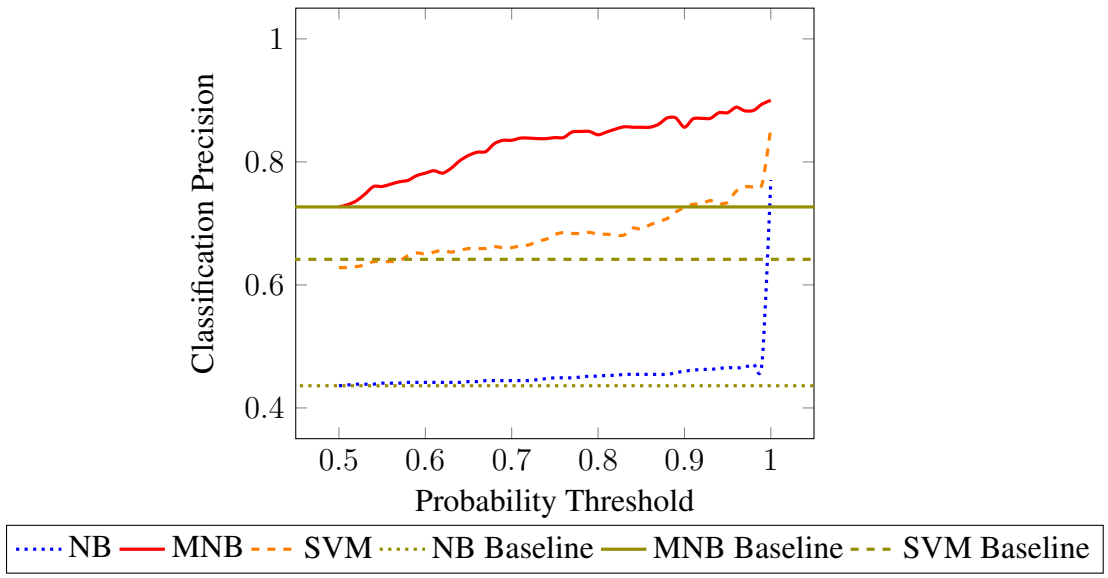


Figure 6.15: Graph of negative class precision results given strong classifier confidence.

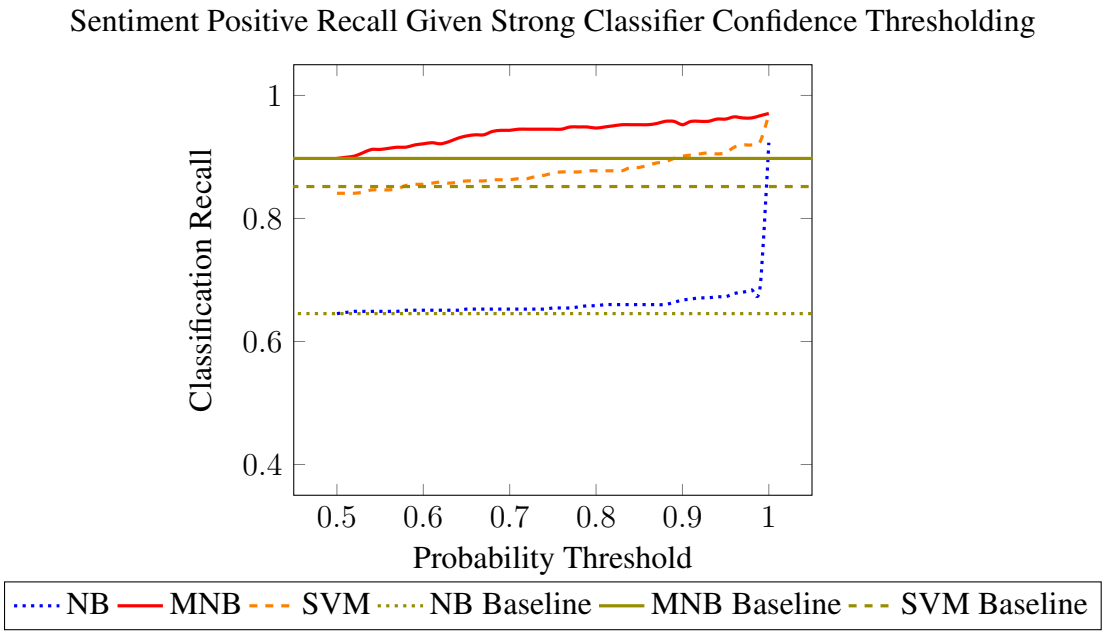


Figure 6.16: Graph of positive class recall results given strong classifier confidence.

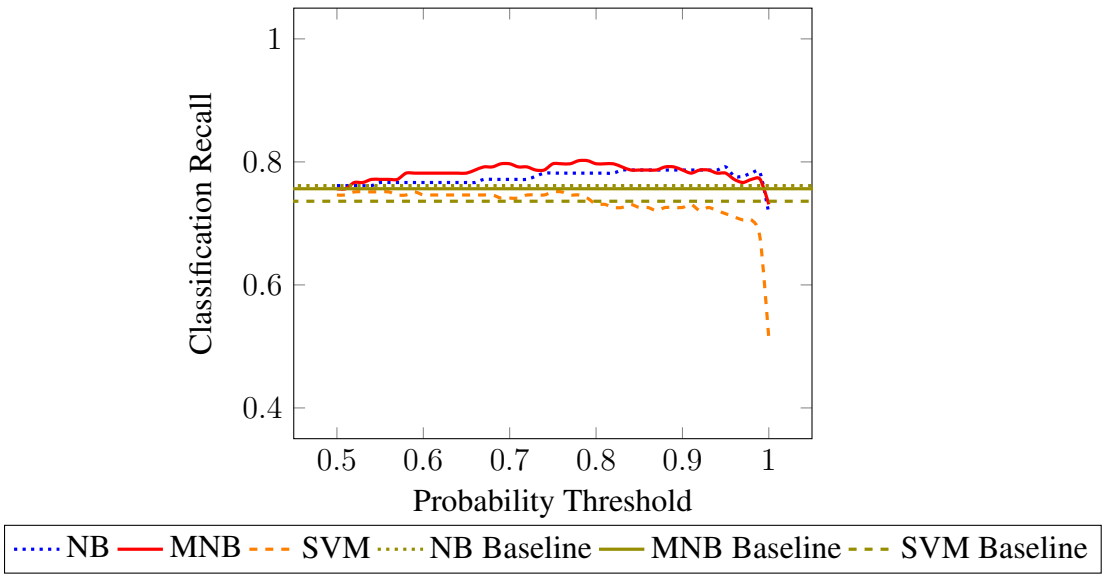


Figure 6.17: Graph of negative class recall results given strong classifier confidence.

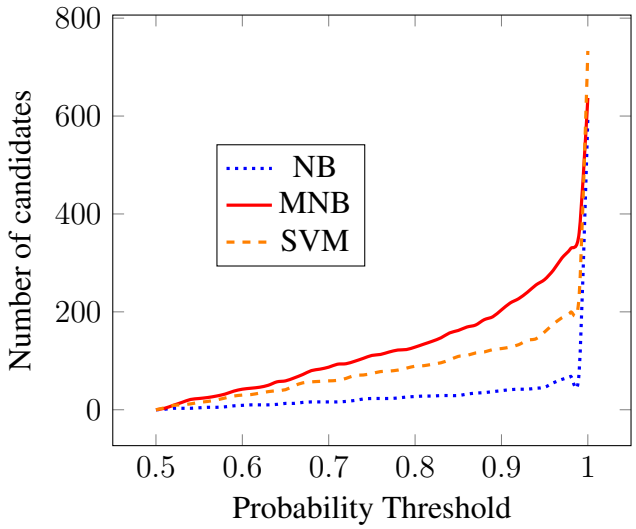


Figure 6.18: Relabelling candidates given varying strong classifier confidence thresholds.

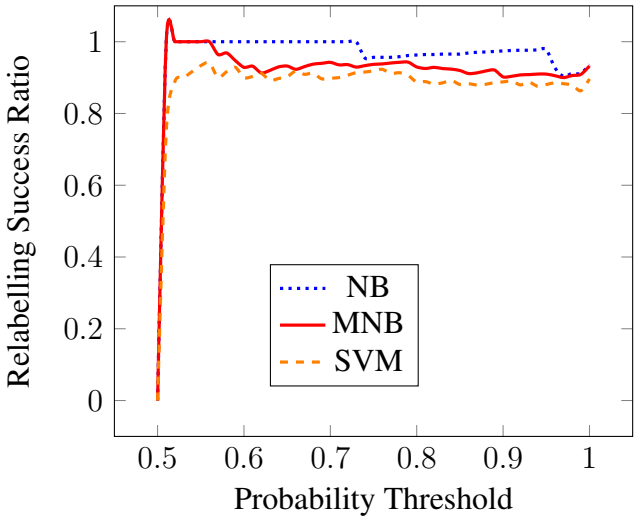


Figure 6.19: Relabelling success rate given varying classifier confidence thresholds.

6.4.4 Results: Document similarity recalibration

The document similarity recalibration protocol was applied to determine whether a level of lexical similarity between a review and its response could be a reliable indicator of shared document sentiment, and hence could be used as a recalibrating factor for sentiment classification. When using the GST algorithm to determine a level of relevance between response and review, it is unclear what value of minimum-matching length (MML) may be appropriate for the task. We initially experiment with two values for the MML: 1 and 2. Accuracy results when using an MML of 2 are shown in figure 6.20, and when using MML1 are shown in figure 6.21. An MML of 1 is the least discriminative approach when using the GST algorithm, which leads to a higher similarity score being produced, and hence a greater range is given when setting this parameter to 1. At an MML of 2, the GST algorithm is slightly more discriminative in its application of the GST algorithm, only producing matching tiles of mutual phrases that are two words or more in length. Common bigrams are less frequent than common unigrams between a response and its review, leading to a smaller range of similarity values. Due to this lack of diversity, experimental results are examined in respect of an MML of 1.

At an MML of 1, accuracy results for all models exceed their respective baselines. At similarity thresholds (the stepped threshold value in Algorithm 5) of 0.01 to 0.02, the recalibration of the NB model causes the results from this classifier to exceed the SVM baseline yielding the accuracies 84.811% and 83.737% respectively. Between threshold value of 0.04 to 0.07, accuracy results for the SVM classifier are within approximately 1% of the MNB baseline. The MNB performs strongly, yielding a maximum accuracy of 88.710% at a similarity of 0.06, before tailing off to the baseline at a similarity of 0.2, and dipping below the baseline slightly, a drop of approximately 1% over the range of similarity scores, between 0.23 and 0.36

The precision and recall trajectories imitate the results given in the previous protocols. Positive precision results show that the MNB exceeds the baseline classification results between the values of 0.03 and 0.36, reaching a maximum precision result for this protocol, 0.924 , at a similarity of 0.1. Recalibration of the NB classifier is beneficial, with results between the threshold values of 0.07 to 0.16 exceeding all baselines and the precision results for the recalibrated MNB

classifier, reach a maximum positive precision of 0.930 at similarity threshold 0.09. The SVM classifier does not yield positive precision results that exceed its baseline, and results tend to this from a starting value of 0.845. The SVM classifier performs somewhat better when examining the negative precision (figure 6.23) and positive recall (figure 6.24) results. In both cases, the SVM model exceeds both the MNB classifier and its baseline, yielding a maximum negative precision of 0.847, at a similarity of 0.01, and a maximum positive recall result of 0.965, also at a similarity of 0.01. The NB classifier does not benefit from recalibration to the extent of the SVM model when observing these metrics, but improvements are made over its baseline between similarity threshold of 0.01 to 0.1. However, the NB model is able to benefit from the results of recalibration when examining the negative recall results (figure 6.25), surpassing the recalibrated MNB results, and achieving a maximum negative recall result of 0.863, between the similarity threshold of 0.09 and 0.12.

The relabelling success rate given the document similarity recalibration protocol is greater than 0.9 for the MNB and NB models, for all similarity thresholds where a level of similarity is detected, and for the SVM model, when the similarity threshold is greater than 0.03 (figure 6.26). When the similarity threshold passes the value of 0.4, all classifiers report near perfect relabelling success. However, there are few candidates for relablelling beyond this threshold, as figure 6.27 shows.

When attempting to reject the null hypothesis that recalibration has no significant effect on the outcome of sentiment classification, the recalibrated SVM and NB classification models using the currently discussed protocol yields statistically significant improvements ($p < 0.01$) over their respective baseline classification models. The NB classifier achieves an accuracy of 84.811% at a similarity of 0.01, and statistical significance remains until a similarity of 0.09. The SVM also achieves statistical significance in the same range of similarity values and achieves a peak accuracy of 85.394% at a similarity score of 0.09. The MNB classifier does not yield statistically significant improvements, as the error rate between the baseline MNB classifier, and the recalibrated classifier is found to yield a comparable number of misclassifications per model.

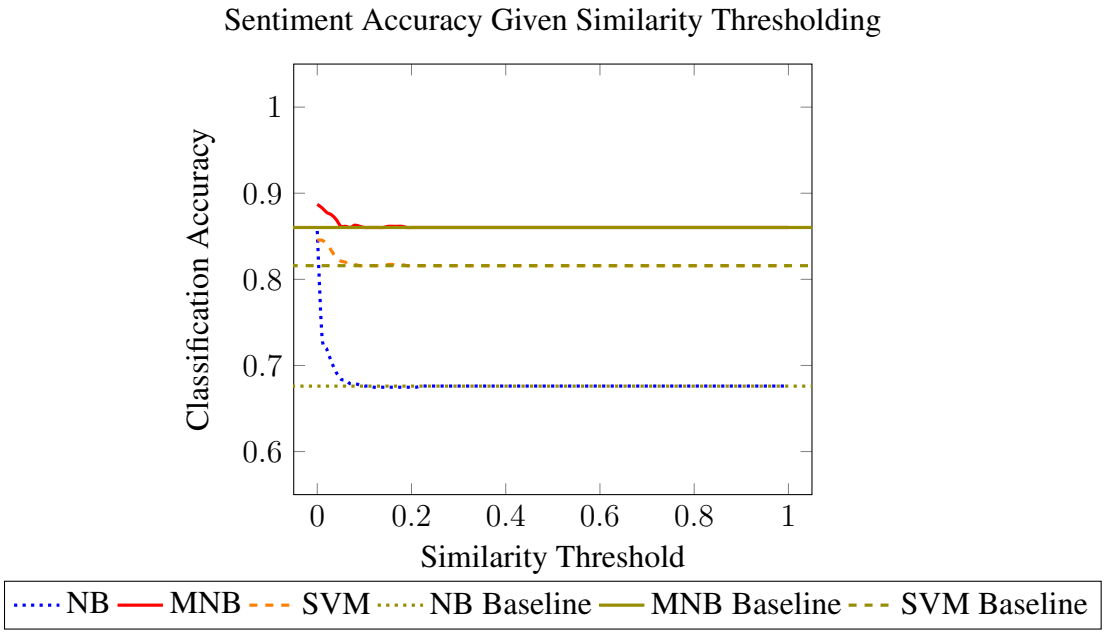


Figure 6.20: Graph of sentiment accuracy results given similarity of review and response (MML = 2).

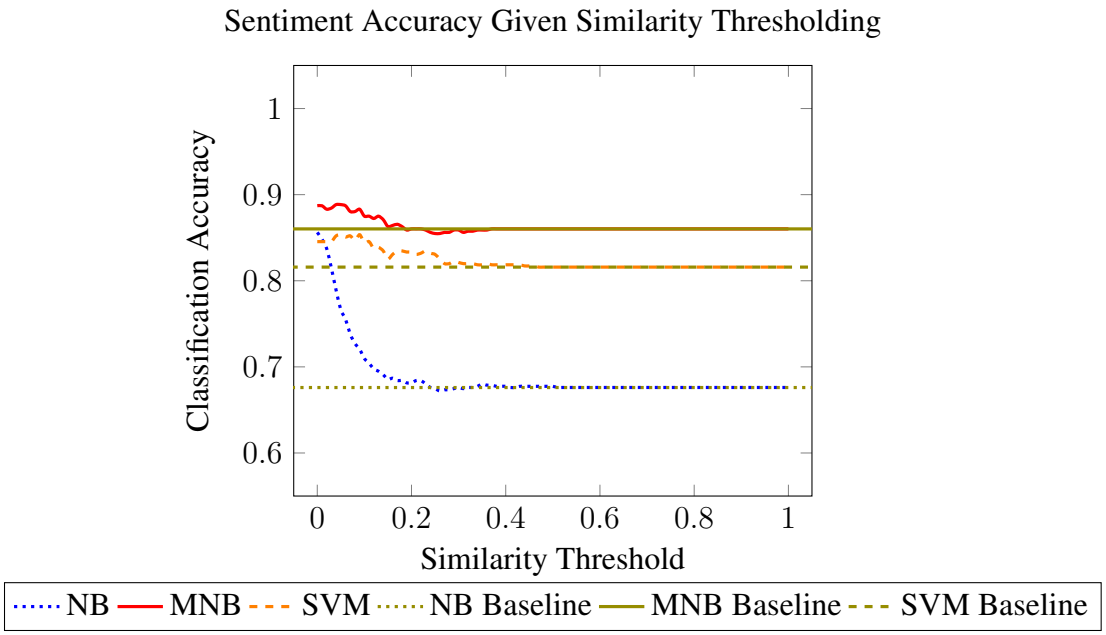


Figure 6.21: Graph of sentiment accuracy results given similarity of review and response (MML = 1).

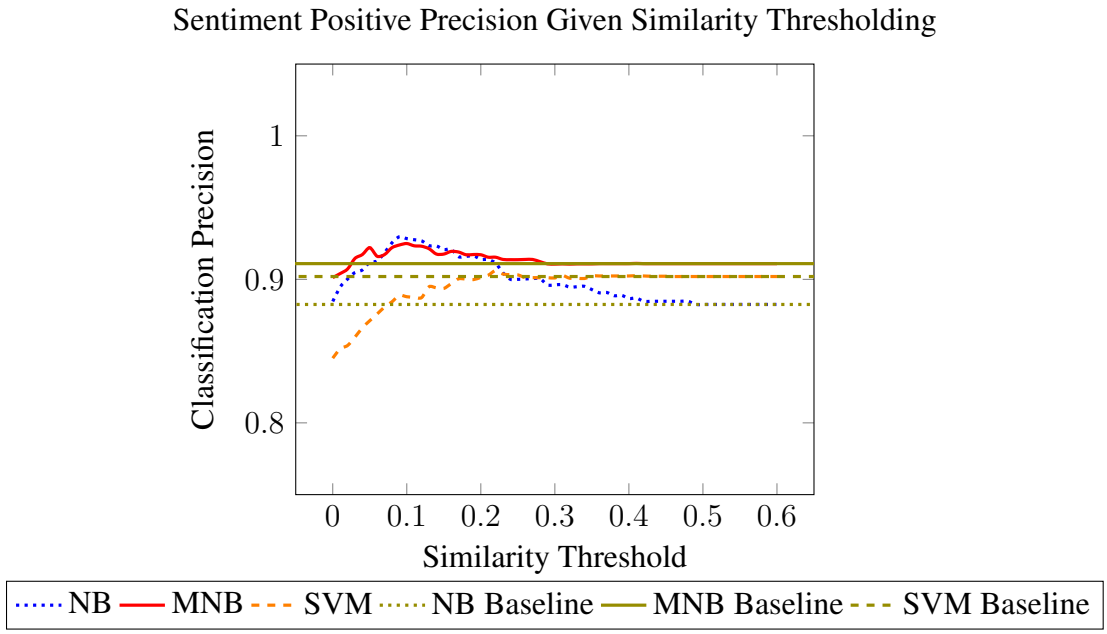


Figure 6.22: Graph of positive class precision results given similarity of review and response (MML=1).

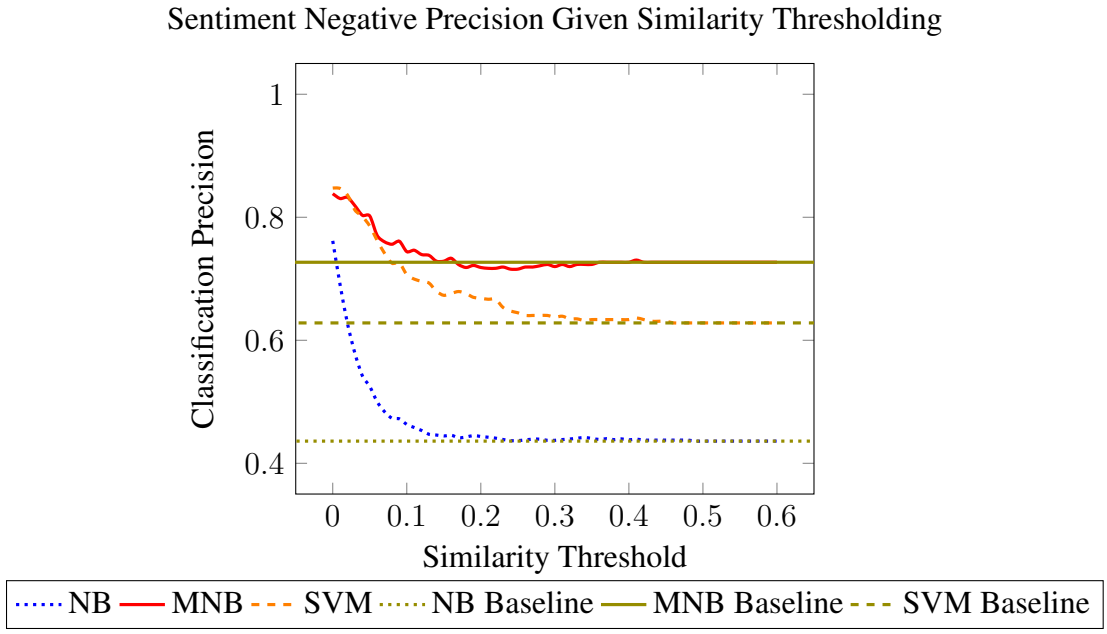


Figure 6.23: Graph of negative class precision results given similarity of review and response (MML=1).

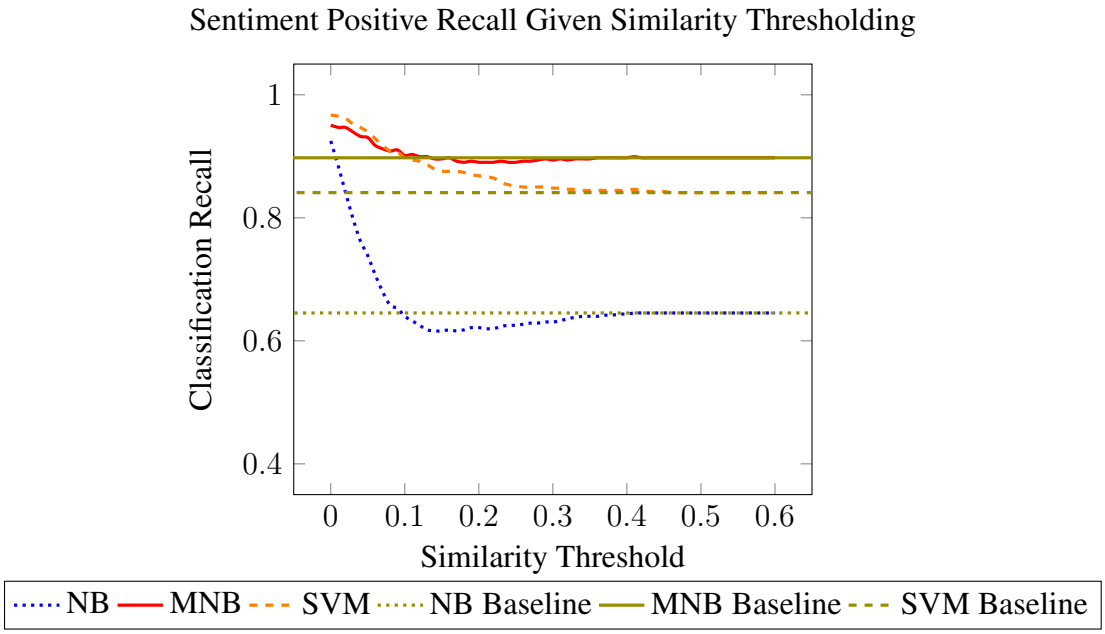


Figure 6.24: Graph of positive class recall results given similarity of review and response (MML=1).

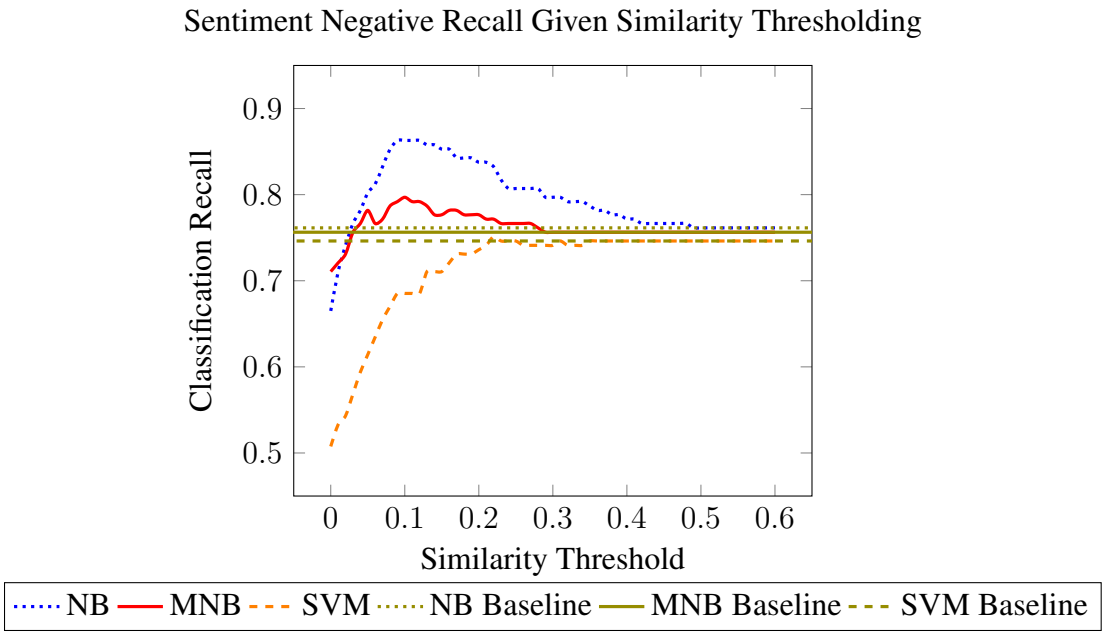


Figure 6.25: Graph of negative class recall results given similarity of review and response (MML=1).

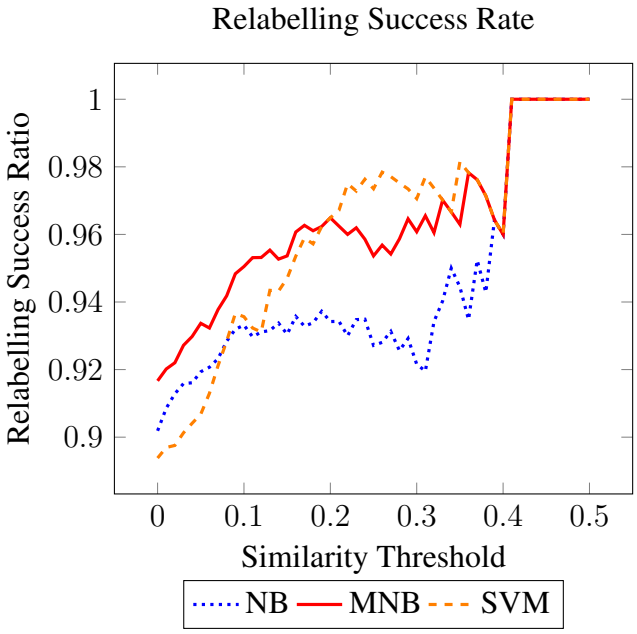


Figure 6.26: Relabelling success rate given varying similarity thresholds (MML=1).

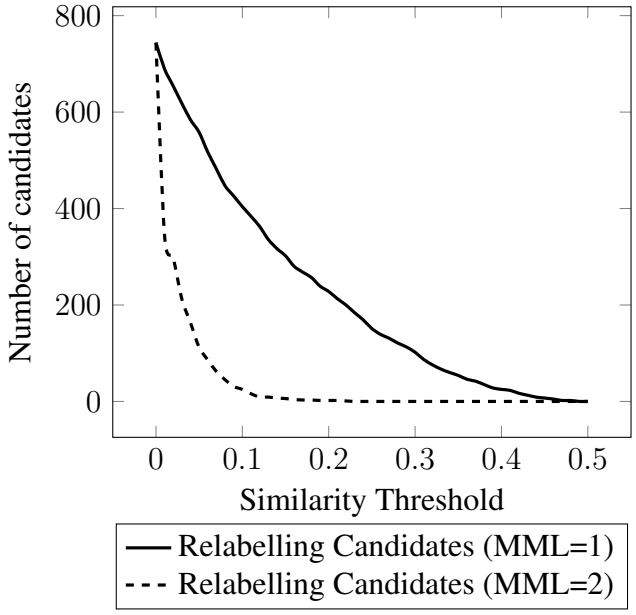


Figure 6.27: Relabelling candidates given varying classifier similarity thresholds. This is model independent for the similarity protocol, hence there is only a single line on this particular graph.

6.5 Discussion

The results across the different protocols demonstrate that the recalibration of classification outcome given the presence of a related document can be achieved, and is successful in yielding positive increases in performance over a baseline method that does not apply any recalibration techniques. The best result is achieved using the first protocol, probabilistic threshold recalibration, using the MNB classification model. Recalibration using this protocol yields a 5.4% increase in accuracy over baseline results, and this improvement over the baseline is found to be statistically significant ($p < 0.01$). This result suggests that for a given sentiment analysis task where a document for classification has a related response, then the probabilistic threshold recalibration would be effective. In particular, a range of threshold values was identified for this particular model to yield statistically significant results over a classification model that doesn't use a recalibration protocol, of 0.54 to 0.99. Particular values in this range, at the probability thresholds of 0.88 and 0.89 were both found to yield the best performance for the MNB model. This would suggest that when attempting to determine what recalibration protocol may be preferable to use, then purely on the basis of the protocol that yields the highest accuracy, then the probabilistic recalibration protocol would be a strong contender.

However, recalibration can be viewed as the process of improving a weak classification model to make it competitive with stronger performing models. In this case, a criterion for selecting the best recalibration protocol may be the protocol that yields the largest improvements in performance. Using this criterion, the document similarity protocol may be deemed an appropriate protocol to yield the best performance due to its ability to increase the outcome of classification for the NB and SVM models. While accuracy results offer a similar increase to the probabilistic thresholding recalibration method, the increases in precision and recall are for these supposed weaker models where the document similarity recalibration protocol shows its strengths, on a class by class basis.

In the results, there are a number of times when results increase dramatically in a single step. This can be seen in figure 6.6, for example, where the NB classification accuracy goes from 71.371% at a threshold of 0.99, to 85.618% when the threshold is 1.0. Also for the

document similarity protocol with the NB classifier, a steep gradient to the results curve can be seen when the accuracy drops from 85.618%, at 0.0, to 75.538%, at a similarity of 0.06. At a probability of 1.0 in the probabilistic threshold recalibration protocol, all response labels are used to recalibrate the outcome of review classification. Similarly in the document similarity protocol, at a threshold of 0.0, where all review-response pairings either contain no computed similarity, or a calculated level of document similarity, then, again, all response labels are used to replace the initial review labellings. In some cases, for example, for the accuracy of the NB classifier, when all response labels replace the review labellings, this is the best accuracy for the NB classifier over all of the increments of the recalibration framework. In this case, the recalibration protocol can be viewed as defunct, as it need not be applied to get this result; all that is required is the replacing of the review labels with the response labels.

In chapter four a number of classification methods were examined that were not subject to the application of the recalibration protocols. In that chapter, classifier choice was examined using a ranking test for significance, the Friedman test. In this test, classifier performance was examined over different review types, and results showed that over the different review types, classifiers performed comparably. In the experiments of this chapter, only a single review type, the type 2 review, is the basis for experimentation. Due to this, the Friedman test would not be able to be suitably applied. The aim of this chapter is to yield improvements in sentiment classification, and improvements can be verified with respect to a baseline. What we see in this chapter, is that in comparison to the baseline methods, all of the recalibration methods are able to yield statistically significant improvements when examined using the McNemar test. The recalibration methods are able to make use of methods based on the notion of recalibration labelling with low classifier confidence, or only recalibrating where there is a level of similarity between response and review. This is a form of correction that standard machine learning methods have not before utilised, as a document set structured like the NCSD has not been examined in this way before. What can be seen is that when recalibration for one classification model is successful, it is on the whole, also successful for the other observed methods. This causes a near consistent ranking of the algorithms, with the MNB as the most successful model,

followed by the SVM and then the NB model. The experiments were carried out with the same classification model acting as both the review and the response classifier. It may be of interest to see if the best performing response classification model, the SVM, could have been used as the response classifier for all experiments, and whether this would yield comparable improvements for all the review classification models.

In the experiments, three recalibration protocols were evaluated that largely fall into either the classes of probabilistic recalibration methods or a document-similarity based method. The latter works at the document text level, and the former works at the classifier confidence levels. While the two approaches are fundamentally different in their approach to recalibration, there is the possibility that the two could complement one another. For example, not only could a label confidence be at a certain threshold, but it could also be stipulated that an element of document similarity could also be required to be present in order for recalibration to occur. However, due to the overall lower performance of the document similarity recalibration protocol, there is the potential that combining methods would not be fruitful, and errors from one recalibration method may propagate when combining both, leading to an overall decrease in performance.

While three methods for classification recalibration given a response were examined, there is also the potential for variations of the currently examined methods to be developed in future work. For example, a static response classification confidence threshold could be imposed which must be surpassed initially before it can be of use for recalibration. Similarly, a classifier confidence threshold could be imposed prior to document similarity classification. Additional protocols could also examine document metadata, such as the length of time between review and response, or the global sentiment of a particular aspect of the health service and whether this aspect was mentioned in either the review or the response.

Finally, comparing our results to the work of Greaves et al. (2013) on the sentiment classification of patient feedback, the best performing method in their experiments, the MNB classifier, achieved an accuracy of 88.6% and an F_1 of 0.89 (figures in their work are only given to 1 and 2 decimal places respectively). Using the probabilistic thresholding recalibration framework, results of our experiments achieve an accuracy of 91.4% and an F_1 of 0.902, yielding state-of-

the-art results for sentiment classification in the patient feedback domain.

Summary

In this chapter, we have examined the role of classifier recalibration for the task of sentiment analysis in patient feedback. The proposed recalibration framework considered acknowledged sentiment in a comment response to recalibrate classifier output. The experimental framework examined three methods for recalibration, two probabilistic and one similarity based. We found that all classifiers exhibited improvements in classification performance when subject to recalibration over varying probability thresholds. Results suggest that the MNB classifier is most suited to the recalibration methods, and yields the best performance, with a 5.4% increase in classification accuracy over the baseline, resulting in an accuracy of 91.4% and F_1 of .902. The proposed recalibration approach is suitable where a dataset contains a number of related documents, similar to a dialogue, and there is the possibility that this method could be expanded in an iterative fashion to discourse data with suitable sentiment annotations.

CHAPTER 7

CONCLUSIONS

Patient feedback forms part of an informal yet vital metric to determine the way in which a health service is handling its care processes and how it is being perceived by the patients that it treats. As the volume of feedback grows due to the increased ease and familiarity with which feedback can be submitted almost instantaneously through web forms and mobile phone applications, there is a need to apply computational methods to deal with the bulk of the data. This data that a health provider receives could be used in many ways: it could be used to change processes in a department, or to inform clinical decisions, or to keep the relevant staff motivated through the knowledge and recognition shown that the care they give is not going unnoticed by those who benefit from it.

This thesis was conceived to examine the application of sentiment analysis to the patient feedback domain. Following extensive experimentation, the work detailed in this thesis has found that the choice of supervised machine learning algorithm or feature representation is not a significant factor in developing an automatic classification system to handle the task. In contrast to this, the type of data that is used to train and test a sentiment classification system for patient feedback was found to have a significant effect on the classification performance and results suggest that in particular, training across review types can be detrimental to a system's overall performance. The study has also sought to determine whether context can be incorporated into the classification process, particularly through the use of the equivalent of annotator rationale, a review response. Recalibration methods developed to incorporate a review's response into the

classification process led to significant improvements in sentiment performance over a baseline method, yielding an accuracy of 91.4%, which while not perfect, exceeds the performance of a number of current systems.

Liu (2012) concludes his study on the topic of sentiment analysis and opinion mining by stating that he does not see a silver bullet solution to sentiment classification occurring in the near future. Despite this, he suggests that work in sentiment classification should examine a large number of diverse application domains in order to contribute gradually to a general solution to the problem. This thesis examines and develops a number of approaches to the classification of sentiment-bearing documents from the patient feedback domain, and in doing so, takes a step towards achieving this goal.

7.1 Contributions

In the following subsections, the contributions of this thesis are recapped in relation to each of the research questions that this thesis sought to answer.

7.1.1 The effects of classifier choice

The empirical work reported in this thesis shows that the choice of supervised machine learning classification model does not have a significant effect on the performance of the sentiment classification of patient feedback. Specifically, the analysis of the five supervised classification models suggested that no single model significantly outperformed another when tested for statistical significance using the Friedman test. However, the Multinomial Naïve Bayes model consistently ranked as one of the best-performing methods for sentiment classification, while the Naïve Bayes and Support Vector Machines approaches were amongst the poorest performing classification models.

Despite no single classifier performing significantly better than any of the other classifiers when compared with a significance test that considered the ranking of each of the classifiers, this result has reassuring implications for those wishing to implement a practical system to anal-

yse the sentiment of patient feedback. The decision about which classification model to choose is often not straightforward, especially where the sentiment classification literature reports differing performance values for the same machine learning models over a variety of different domains. Instead, classifier selection may be made with more practical constraints in mind, such as speed or memory requirements. In this case, the Multinomial Naïve Bayes may be preferable not only for its consistently high ranking performance, but also because over a single run of the experiments in this thesis, the time to train the model was 0.07 seconds on a 1.3 GHz Intel Core i5 processor with 4GB of DDR3 memory, significantly faster than training a Support Vector Machine, which takes 4.41 seconds on the same machine.

7.1.2 The effects of feature choice

The empirical work in this thesis examined this research question by examining the application of nine different types of feature to determine the effect of feature choice on the outcome of sentiment classification in the clinical domain. It was found that of the feature variations examined using the best performing supervised machine learning model, the Multinomial Naïve Bayes classifier, that there was no significant difference in the outcome of patient feedback classification over the examined features when using the Friedman test ($p < 0.05$). Again, this result is reassuring and indicates that the choice of feature could be left to the developer of a given system for the sentiment classification of patient feedback. The choice could instead be left to more practical issues, such as the time to convert the document into a document vector using the chosen feature representation.

This work examined different text pre-processing and term weighting methods. While no significant differences were found in the outcome of sentiment classification using these different feature representations, some were found to consistently rank better than others when tested on different review types. For example, for Type 1 reviews, features that were lower-case with a boolean weighting ranked as the best approach, whereas lower-casing with stop words removed ranked as the representation that yielded the poorest results. The superior performance of the boolean weighting confirms the assumptions outlined by Pang et al. (2002) that term presence

information was preferable to term frequency information when using machine learning classifiers to categorise movie reviews by sentiment, and our experiments have confirmed that this also generalises to the sentiment classification of patient feedback. This result is also indicative of the positive effect that stopwords have upon sentiment classification, and suggests that sentiment classification requires the information contained in stopwords, so it may not be advisable to remove these, to yield increases in classification performance. For Type 2 reviews, the TF-IDF term weighting yielded the best results, closely followed by a lower-case word stem representation. It should be noted that each of the feature representations was examined independently, but there is the possibility in this instance that the preprocessing and term weighting scheme could be combined to yield a combined best performance. Finally, for Type 3 reviews, term weights that were normalised by document length yielded the best performance, again closely followed by the conversion of the text to lower-case stemmed words. Wordcount term weights yielded the poorest results when examining this research question in respect of the Type 3 reviews, again confirming the assumptions made by Pang et al. (2002).

7.1.3 The effects of review type

The NCSD contained patient feedback in three review formats: likes and dislikes (Type 1), advice (Type 2), and a combination of all three (Type 3). Initial experiments found that the choice of review type does not significantly affect the outcome of sentiment classification when training and testing across the same review type. However, training and testing across review type, especially from Type 2 to Type 1 does have significant effects on machine learning models, to the detriment of the performance of sentiment classification.

This was one of the few experiments whereby at a significance of $\alpha = 0.05$ a significant difference in classification results was found between the performance on Type 1 data alone and the cross-testing of training on Type 2 and testing on Type 1. This would suggest that the type of data that is used for sentiment classification is an important issue to consider, especially when training on Type 2 reviews and testing on the Type 1 reviews. Interestingly, training on the Type 1 and testing on Type 2 reviews, although demonstrating relatively poor performance,

was not found to be significantly detrimental to the overall outcome of sentiment classification. Due to this, systems should be wary of the type of data used to train machine learning classifiers for the classification of patient feedback by conveyed sentiment, and if possible, train on data that is of the same type that the system will be applied to.

When considering review type in classification, the use of the final sentences of a review for classification purposes was also examined. A final sentence summarises a review, and in doing so also tends to summarise the overall sentiment of the review. While not being as informative in regards to the content of the review, a test of the best performing classifier was able to yield an accuracy of 81.33% on the final sentences of type two reviews. This is competitive with the 84.07% accuracy that was gained when classifying the whole review, considering that a document length a fraction of the size of the whole is able to be used to classify the whole document's sentiment so accurately. Where very large datasets are to be classified, such a technique may be suitable in scaling the input documents to a size that enables classification in a time that is practical to the requirements of the user of such software.

7.1.4 Classifier recalibration

The results of the supervised machine learning experiments were encouraging, and the classifiers generally performed well. However, a misclassification analysis found that issues with the training data, spelling errors and the implicit communication of sentiment all resulted in errors occurring during sentiment classification. While the first two issues could potentially have been solved through the use of more training data and a spell checker, the detection and correct classification of implicit instances of sentiment conveyance is not as straightforward a problem to solve. Due to this, a recalibration framework was proposed to consider the context offered by a response when determining the sentiment of a review, in order to gauge if sentiment was present in a review where a supervised classification model struggled.

The framework incrementally investigated three recalibration protocols: probabilistic threshold recalibration, strong probabilistic threshold recalibration and document similarity recalibration. The probabilistic thresholding protocol was based on the notion that if the review labelling

confidence was low, there was a possibility of doubt in the proposed labelling, and a process of recalibrating with a response label may either confirm that the original label was accurate, or correct the labelling where the classification model has incorrectly classified the review, irrespective of the review label confidence. The strong probabilistic recalibration relabelled the review with its own only if the response label confidence was equal to or greater than that of the review. This constraint resulted in the examined classifiers being more selective in forming the set of candidates for recalibration, and improvements were not yielded beyond those offered by the first protocol, the probabilistic thresholding recalibration protocol. The final recalibration protocol was based upon the notion of document similarity, between the review and its response. This used the response label for recalibration only if a level of similarity between the review and response was surpassed. Typically similarity was low, meaning that classification outcome was only recalibrated when document similarity was between a score of 0.01 and 0.28. Significant increases over the baseline were yielded for all machine learning methods when similarity was considered. However, this method performed no better than a blanket recalibration of review labellings, irrespective of the similarity between a review and its response.

Overall, review responses were found to be useful in significantly improving the classification outcome of a collection of patient feedback by the sentiment conveyed by each of the reviews. Results of the recalibration framework yielded a peak accuracy of 91.4% using the probabilistic threshold recalibration protocol with the Multinomial Naïve Bayes model at a threshold value of 0.88. This is a significant increase in performance over a baseline classification approach ($p < 0.01$), which highlights the benefits of the application of the recalibration protocol where classifier confidence was not at its maximum. While this question focused on the use of a review response labelling as a recalibrating factor for patient feedback classification, we believe there is the potential for this to be extended to other domains for sentiment analysis where a review has one or a number of related documents that could be used for recalibration to improve the overall performance of the sentiment classification system.

7.2 Future work

7.2.1 Emotion classification of patient feedback

The work in this thesis has examined techniques for the automatic classification of patient feedback as either conveying a positive or negative sentiment. While these categories of classification are adequate for gleaning basic evaluative information about the content of a review, the spectrum of human evaluation goes beyond the binary distinction of sentiment that we have used in this thesis to span a range of emotions, each of which has the potential to be conveyed through a text (Ortony et al., 1988). Due to this, there is the possibility that the task of sentiment analysis that has been the focus of this thesis can be expanded to the classification of emotion that patient feedback conveys.

When considering other domains in the sentiment classification literature, such as film reviews, for example, it could be argued that the evaluation of emotion is not such an important issue, aside from perhaps determining the emotions invoked by a particular film. Due to the personal nature of healthcare, we could assume that an item of patient feedback could potentially be more emotive than a movie review in what it conveys through its content. For example, by examining and classifying instances of emotion in patient feedback, for content that conveys perhaps anger, disgust or sadness, it may be possible for a health service to monitor and react in a more sensitive way to the deeper insights revealed by determining such an emotion that has been articulated by a reviewer. Additionally, a health service could examine the aspects of patient care that are making a patient happy or joyful, and what amidst the care processes offered to them is causing surprise, or perhaps even fear.

Coinciding with the work of this thesis, we developed a system for the automatic classification of emotions in news headline data (Smith & Lee, 2012). In this work, the headlines were annotated with information denoting the strength of six potential emotions conveyed through the text. There is the potential in future work stemming from this thesis for the NCSD to be annotated with a similar labelling scheme. This would potentially extend the current labelling scheme, and enable the suite of supervised machine learning models to be re-evaluated within

the scope of a more detailed annotation scheme, and the question of whether emotional analysis of patient feedback can be robustly performed could be examined.

By extending this work to classify the emotion in patient feedback, interesting insights could also be revealed in relation to the recalibration framework. The recalibration framework relies on a two-way interaction between a patient and the health service, but only so far as recalibrating where a classification system may not have initially classified the sentiment of a review correctly. By annotating both the feedback and the response with potentially conveyed emotions, the recalibration framework would have to adapt to the norms of emotive interaction. For example, while unhappiness in patient feedback may be responded to with sorrow by a health service provider, and happiness with joy, fear or anger may be responded to in a number of different ways. Determining an appropriate taxonomy of emotions that fitted those conveyed by patient feedback and the response to patient feedback, and the interactions between the two parties would be important in such an extension of the current work.

7.2.2 Deep learning for the sentiment classification of patient feedback

The approach to the sentiment classification of patient feedback in this thesis examined the application of supervised machine learning models trained for the task. However, the given approaches modelled language as a bag of words in the machine learning process, and hence details regarding syntax and semantics could possibly have been lost. The concept of deep learning has been proposed as a solution to accurately learning the kind of functions represented by high-level abstractions, such as languages (Bengio, 2009). The deep learning approach attempts to model multiple levels of non-linear operations through the construction of a deep architecture. One example of such a deep architecture is a neural network with several hidden layers, with each modelling a particular aspect of a machine learning problem.

Deep learning has been successfully applied to natural language processing tasks such as machine translation (Luong et al., 2015), but in relation to its application to sentiment analysis it has produced modest improvements in classification when classifying the sentiment of film reviews (Socher et al., 2013) using a deep learning technique known as a recursive neural tensor

network. This technique yielded a 5% increase in accuracy when compared to an NB classifier trained for the same task, and 3% in comparison to an SVM model, and hence the application of a deep learning method should yield comparable increases in the classification of patient feedback by sentiment.

Additionally, and perhaps more interestingly, deep learning could be beneficial to the recalibration protocols due to the application of the deep learning architecture to model the recalibrating factor of a response. The recalibration framework could be developed further to incorporate a deep learning architecture that not only considers the conveyance of sentiment in an isolated document, but also the behaviour of a response in the hidden layers of a neural network-based model.

7.2.3 Domain adaptation

This thesis focused on the sentiment analysis of patient feedback, but by no means are the methods developed limited to only this domain. The recalibration framework was developed to improve the standard of classification where a review potentially had related, context-bearing documents associated with it, and patient feedback is not the only source of data with such a structure. In the sphere of online reviews, increasingly, reviews are not isolated from responses, and so, there are a number of potential domains where the methods developed in this thesis could be applied. Social media sites such as Facebook, Twitter and Google+ all enable businesses to have their own form of online presence, and in doing so, allow users to post reviews, but also allow the businesses to post responses to the reviews. A suitable web crawler could be set up to download data from a number of different domains, and in turn, the recalibration framework could be examined on the crawled data.

Aside from social media, sites such as TripAdvisor have also implemented a similar feedback mechanism to the one used by NHS Choices; namely, where a review is given about a particular service provider, a representative from that organisation is able to respond to the given review. In the case of TripAdvisor, this gives the particular business owner the opportunity to respond appropriately to a review, perhaps thanking a customer and clarifying any issues

that a customer may have had. This two-way interaction can be integrated into the recalibration framework developed in this thesis, and could potentially produce a better standard of sentiment classification on data from the hotel and restaurant review domain. For example, Proserpio & Zervas (2014) study the impact of management responses on a company's online reputation. Their study focuses on the use of TripAdvisor data for this purpose, and they collect a number of hotel reviews and responses. They did not use any sentiment analysis techniques for the given study, so future work could be extended to work on data such as theirs, to examine whether the recalibration framework is able to produce state-of-the-art results on the hotel and restaurant review domains also.

Appendices

LIST OF REFERENCES

Abdul-Mageed, M., Diab, M. T. & Korayem, M. (2011), "Subjectivity and Sentiment Analysis of Modern Standard Arabic", *in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA: ACL, pp. 587–591.

Agrawal, R., Rajagopalan, S., Srikant, R. & Xu, Y. (2003), "Mining Newsgroups Using Networks Arising from Social Behavior", *in Proceedings of the 12th International World Wide Web Conference*, New York, NY, USA: ACM Press, pp. 529–535.

Akkaya, C., Wiebe, J. & Mihalcea, R. (2009), "Subjectivity Word Sense Disambiguation", *in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore: ACL, pp. 190–199.

Alhessi, Y. & Wicentowski, R. (2015), "SWATAC: A Sentiment Analyzer using One-Vs-Rest Logistic Regression", *in Proceedings of the 9th International Workshop on Semantic Evaluation*, Denver, Colorado, USA: ACL, pp. 636–639.

Ali, T., Schramm, D., Sokolova, M. & Inkpen, D. (2013), "Can I Hear You? Sentiment Analysis on Medical Forums", *in Sixth International Joint Conference on Natural Language Processing*, Nagoya, Japan: AFNLP / ACL, pp. 667–673.

Andreevskaia, A. & Bergler, S. (2008), "When Specialists and Generalists Work Together: Overcoming Domain Dependence in Sentiment Tagging", *in Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, USA: ACL, pp. 290–298.

Anthony, L. (2011), *AntConc (Version 3.2.4m) [Computer Software]*, Waseda University, Tokyo, Japan, available from `http://www.laurenceanthony.net/`.

Anthony, L. (2015), *TagAnt (Version 1.2.0) [Computer Software]*, Waseda University, Tokyo, Japan, available from `http://www.laurenceanthony.net`.

Artstein, R. & Poesio, M. (2008), "Inter-coder Agreement for Computational Linguistics", *Computational Linguistics* 34(4), pp. 555–596.

Asher, N., Benamara, F. & Mathieu, Y. Y. (2008), "Distilling Opinion in Discourse: A Preliminary Study", *in Proceedings of the 22nd International Conference on Computational Linguistics*, Manchester, UK: Coling 2008 Organising Committee, pp. 7–10.

Aue, A. & Gamon, M. (2005), "Customizing Sentiment Classifiers to New Domains: A Case Study", *in Proceedings of 5th International Conference on the Recent Advances in Natural Language Processing*, Borovets, Bulgaria: RANLP Organising Committee.

Austin, J. (1962), *How to Do Things With Words*, Oxford University Press.

Baccianella, S., Esuli, A. & Sebastiani, F. (2010), "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining", *in Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Valletta, Malta: ELRA, pp. 2200–2204.

Baker, P., Hardie, A. & McEnery, T. (2006), *A Glossary of Corpus Linguistics*, Edinburgh University Press.

Balahur, A., Hermida, J. M. & Montoyo, A. (2011), "Detecting Implicit Expressions of Sentiment in Text Based on Commonsense Knowledge", *in Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, Portland, Oregon: ACL, pp. 53–60.

Becker, I. & Aharonson, V. (2010), "Last but Definitely Not Least: On the Role of the Last Sentence in Automatic Polarity-Classification", *in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: Short Papers*, Uppsala, Sweden: ACL, pp. 331–335.

Beineke, P., Hastie, T., Manning, C. & Vaithyanathan, S. (2003), "An exploration of sentiment summarization", *in Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*, volume 3, pp. 12–15.

Bengio, Y. (2009), "Learning Deep Architectures for AI", *Foundations and Trends in Machine Learning* 2(1), pp. 1–127.

Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V. & Jurafsky, D. (2006), "Extracting Opinion Propositions and Opinion Holders using Syntactic and Lexical Cues", *in Computing Attitude and Affect in Text: Theory and Applications*, Springer, pp. 125–141.

Biber, D. (2009), "A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing", *International Journal of Corpus Linguistics* 14(3), pp. 275–311.

Bilgic, M., Namata, G. & Getoor, L. (2007), "Combining Collective Classification and Link Prediction", *in Workshops Proceedings of the 7th IEEE International Conference on Data Mining*, Omaha, Nebraska, USA: IEEE Computer Society, pp. 381–386.

Biyani, P., Caragea, C., Mitra, P., Zhou, C., Yen, J., Greer, G. E. & Portier, K. (2013), "Co-training over Domain-independent and Domain-dependent Features for Sentiment Analysis of an Online Cancer Support Community", *in Proceedings of the IEEE Conference on Advances in Social Networks Analysis and Mining*, Niagara, ON, Canada: ACM Press, pp. 413–417.

Blitzer, J., Dredze, M. & Pereira, F. (2007), "Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification", *in Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic: ACL, pp. 440–447.

Bloom, K., Garg, N. & Argamon, S. (2007), "Extracting Appraisal Expressions", *in Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, Rochester, New York, USA: ACL, pp. 308–315.

Bobicev, V., Sokolova, M., Jafer, Y. & Schramm, D. (2012), "Learning Sentiments from Tweets with Personal Health Information", *in Proceedings of the 25th Canadian Conference on Artificial Intelligence*, Toronto, ON, Canada,: Springer, pp. 37–48.

Bollegala, D., Weir, D. & Carroll, J. (2011), "Using Multiple Sources to Construct a Sentiment Sensitive Thesaurus for Cross-Domain Sentiment Classification", *in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Techologies*, ACL, pp. 132–141.

Bollen, J., Mao, H. & Zeng, X. (2011), "Twitter mood predicts the stock market", *Journal of Computational Science* 2(1), pp. 1–8.

Breiman, L. (1996), "Bagging predictors", *Machine Learning* 24(2), pp. 123–140.

Breiman, L. (2001), "Random Forests", *Machine Learning* 45(1), pp. 5–32.

Buckland, M. K. & Gey, F. C. (1994), "The relationship between recall and precision", *Journal of the American Society for Information Science* 45(1), pp. 12–19.

Buckley, C. (1985), "Implementation of the SMART information retrieval system", Technical report, Cornell University.

Cambria, E., Benson, T., Eckl, C. & Hussain, A. (2012), "Sentic PROMs: Application of sentic computing to the development of a novel unified framework for measuring health-care quality", *Expert Systems with Applications* 39(12), pp. 10533 – 10543.

Carletta, J. (1996), "Assessing Agreement on Classification Tasks: The Kappa Statistic", *Computational Linguistics* 22(2), pp. 249–254.

Carvalho, P., Sarmento, L., Teixeira, J. & Silva, M. J. (2011), "Liars and Saviors in a Sentiment Annotated Corpus of Comments to Political Debates", *in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA: ACL, pp. 564–568.

Chen, Y. & Skiena, S. (2014), "Building Sentiment Lexicons for All Major Languages", *in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, MD, USA: ACL, pp. 383–389.

Choi, Y. & Cardie, C. (2009), "Adapting a Polarity Lexicon using Integer Linear Programming for Domain-Specific Sentiment Classification", *in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore: ACL, pp. 590–598.

Chong, M. & Specia, L. (2012), "Linguistic and Statistical Traits Characterising Plagiarism", *in Proceedings of the 24th International Conference on Computational Linguistics*, Mumbai, India: Indian Institute of Technology Bombay, pp. 195–204.

Chung, J. E. & Mustafaraj, E. (2011), "Can Collective Sentiment Expressed on Twitter Predict Political Elections?", *in Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, San Francisco, California, USA: AAAI Press, pp. 1770–1771.

Cieliebak, M., Dürr, O. & Uzdilli, F. (2014), "Meta-Classifiers Easily Improve Commercial Sentiment Detection Tools", *in Proceedings of the 9th edition of the Language Resources and Evaluation Conference*, pp. 3100–3104.

Ciresan, D. C., Meier, U., Gambardella, L. M. & Schmidhuber, J. (2010), "Deep, Big, Simple Neural Nets for Handwritten Digit Recognition", *Neural Computation* 22(12), pp. 3207–3220.

Clough, P. (2003), *Measuring Text Reuse*, Ph.D. thesis, Department of Computer Science, University of Sheffield.

Clough, P., Gaizauskas, R., Piao, S. S. & Wilks, Y. (2002), "METER: MEasuring TExt Reuse", *in Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA: ACL, pp. 152–159.

Cohen, J. (1960), "A Coefficient of Agreement for Nominal Scales", *Educational and Psychological Measurement* 20, pp. 37–46.

Dai, L., Chen, H. & Li, X. (2011), "Improving sentiment classification using feature highlighting and feature bagging", *in 11th IEEE International Conference on Data Mining Workshops*, Vancouver, BC, Canada: IEEE, pp. 61–66.

Dave, K., Lawrence, S. & Pennock, D. M. (2003), "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews", *in Proceedings of the 12th International World Wide Web Conference*, Budapest, Hungary: ACM, pp. 519–528.

de Kauter, M. V., Desmet, B. & Hoste, V. (2015), "The good, the bad and the implicit: a comprehensive approach to annotating explicit and implicit sentiment", *Language Resources and Evaluation* 49(3), pp. 685–720.

De Smedt, T. & Daelemans, W. (2012), "" Vreselijk mooi!"(terribly beautiful): A Subjectivity Lexicon for Dutch Adjectives.", *in Proceedings of the Eighth International Conference on Language Resources and Evaluation*, Istanbul, Turkey: ELRA, pp. 3568–3572.

Demšar, J. (2006), "Statistical comparisons of classifiers over multiple data sets", *The Journal of Machine Learning Research* 7, pp. 1–30.

Demšar, J., Curk, T., Erjavec, A., Črt Gorup, Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M. & Zupan, B. (2013), "Orange: Data Mining Toolbox in Python", *Journal of Machine Learning Research* 14, pp. 2349–2353.

Denecke, K. & Deng, Y. (2015), "Sentiment analysis in medical settings: New opportunities and challenges", *Artificial Intelligence in Medicine* 64, pp. 17–27.

Deng, Y., Stoehr, M. & Denecke, K. (2014), "Retrieving attitudes: Sentiment analysis from clinical narratives", *in Proceedings of the Medical Information Retrieval Workshop at SIGIR 2014*, Gold Coast, Australia: CEUR-WS.org, pp. 12–15.

Dermouche, M., Khouas, L., Velcin, J. & Loudcher, S. (2013), "AMI&ERIC: How to Learn with Naive Bayes and Prior Knowledge: an Application to Sentiment Analysis", *in Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, USA: ACL, pp. 364–368.

Devitt, A. & Ahmad, K. (2007), "Sentiment Polarity Identification in Financial News: A Cohesion-based Approach", *in Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic: ACL, pp. 984–991.

Dietterich, T. G. (1998), "Approximate statistical tests for comparing supervised classification learning algorithms", *Neural computation* 10(7), pp. 1895–1923.

Ding, X., Liu, B. & Yu, P. S. (2008), "A Holistic Lexicon-based Approach to Opinion Mining", *in Proceedings of the 2008 International Conference on Web Search and Data Mining*, New York, NY, USA: ACM Press, pp. 231–240.

dos Santos, C. N. & Gatti, M. (2014), "Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts", *in Proceedings of the 25th International Conference on Computational Linguistics*, Dublin, Ireland: ACL, pp. 69–78.

Duan, W., Gu, B. & Whinston, A. B. (2008), "Do online reviews matter? - An empirical investigation of panel data", *Decision Support Systems* 45(4), pp. 1007–1016.

Dubout, C. & Fleuret, F. (2014), "Adaptive sampling for large scale boosting", *Journal of Machine Learning Research* 15(1), pp. 1431–1453.

Dunning, T. (1993), "Accurate Methods for the Statistics of Surprise and Coincidence", *Computational Linguistics* 19(1), pp. 61–74.

Engström, C. (2004), *Topic Dependence in Sentiment Classification*, Master's thesis, St. Edmunds College, University of Cambridge.

Eskander, R. & Rambow, O. (2015), "SLSA: A Sentiment Lexicon for Standard Arabic", *in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal: ACL, pp. 2545–2550.

Esuli, A. & Sebastiani, F. (2005), "Determining the semantic orientation of terms through gloss classification", *in Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management*, Bremen, Germany: ACM Press, pp. 617–624.

Esuli, A. & Sebastiani, F. (2006), "SENTIWORDNET: A publicly available lexical resource for opinion mining", *in Proceedings of the 5th Conference on Language Resources and Evaluation*, Genova, Italy: ELRA, pp. 417–422.

Feng, S., Kang, J. S., Kuznetsova, P. & Choi, Y. (2013), "Connotation Lexicon: A Dash of Sentiment Beneath the Surface Meaning", *in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria: ACL, pp. 1774–1784.

Firth, J. R. (1957), *Studies in Linguistic Analysis*, Basil Blackwell, Oxford, chapter A Synopsis of Linguistic Theory 1930-1955, pp. 1–32.

Fisher, R. A. (1932), *Statistical Methods for Research Workers*, Edinburgh: Oliver & Boyd.

Friedman, M. (1937), "The use of ranks to avoid the assumption of normality implicit in the analysis of variance", *Journal of the American Statistical Association* 32(200), pp. 675–701.

Gamallo, P. & García, M. (2014), "Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets", *in Proceedings of the 8th International Workshop on Semantic Evaluation*, Dublin, Ireland: ACL, pp. 171–175.

Ganu, G., Elhadad, N. & Marian, A. (2009), "Beyond the Stars: Improving Rating Predictions using Review Text Content", *in Proceedings of the 12th International Workshop on the Web and Databases*, Providence, Rhode Island, USA: WebDB, pp. 1–6.

Georgiou, D., MacFarlane, A. & Tony, R.-R. (2015), "Extracting sentiment from healthcare survey data: An evaluation of sentiment analysis tools", *in Proceedings of the Science and Information Conference 2015*, London: IEEE, pp. 352–361.

Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J. A. & Reyes, A. (2015), "SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter", *in Proceedings of the 9th International Workshop on Semantic Evaluation,*, Denver, Colorado, USA: ACL, pp. 470–478.

Ginzburg, J. (2010), "Relevance for Dialogue", *in* Łupkowski, P. & Purver, M. (Eds.), *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue*, Poznań: Polish Society for Cognitive Science, pp. 121–129.

Glorot, X., Bordes, A. & Bengio, Y. (2011), "Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach", *in Proceedings of the 28th International Conference on Machine Learning , June 28 - July 2, 2011*, Bellevue, Washington, USA: Omnipress, pp. 513–520.

Goeuriot, L., Na, J.-C., Min Kyaing, W. Y., Khoo, C., Chang, Y.-K., Theng, Y.-L. & Kim, J.-J. (2012), "Sentiment Lexicons for Health-related Opinion Mining", *in Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, New York, NY, USA: ACM Press, pp. 219–226.

Goldstein, J., Kantrowitz, M., Mittal, V. & Carbonell, J. (1999), "Summarizing Text Documents: Sentence Selection and Evaluation Metrics", *in Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, CA, USA: ACM, pp. 121–128.

González-Ibáñez, R., Muresan, S. & Wacholder, N. (2011), "Identifying Sarcasm in Twitter: A Closer Look", *in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA: ACL, pp. 581–586.

GOV.UK (2012), "Friends and family test aims to improve patient care and identify best performing hospitals", `http://tinyurl.com/friend-family-2012`, last accessed: 24-11-2015.

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A. & Donaldson, L. (2013), "Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online", *Journal of Medical Internet Research* 15(11).

Greene, S. & Resnik, P. (2009), "More than Words: Syntactic Packaging and Implicit Sentiment", *in Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Boulder, Colorado: ACL, pp. 503–511.

Grice, H. P. (1970), *Syntax and Semantics*, Academic Press, volume 3: Speech Acts, chapter Logic and Conversation.

Hagen, M., Potthast, M., Büchner, M. & Stein, B. (2015), "Webis: An Ensemble for Twitter Sentiment Detection", *in Proceedings of the 9th International Workshop on Semantic Evaluation*, Denver, Colorado, USA: ACL, pp. 582–589.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. (2009), "The WEKA data mining software: an update", *SIGKDD Explorations* 11(1), pp. 10–18.

Hamdan, H., Bellot, P. & Béchet, F. (2015), "Lsislif: CRF and Logistic Regression for Opinion Target Extraction and Sentiment Polarity Analysis", *in Proceedings of the 9th International Workshop on Semantic Evaluation*, Denver, Colorado, USA: ACL, pp. 753–758.

Hatzivassiloglou, V. & McKeown, K. (1997), "Predicting the Semantic Orientation of Adjectives", *in Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, Spain: ACL, pp. 174–181.

He, Y., Lin, C. & Alani, H. (2011), "Automatically Extracting Polarity-Bearing Topics for Cross-Domain Sentiment Classification", *in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologie*, Portland, Oregon, USA: ACL, pp. 123–131.

Heider, F. (1946), "Attitudes and cognitive organization", *Journal of Psychology* 21(58), pp. 107–112.

Hopper, A. M. & Uriyo, M. (2015), "Using sentiment analysis to review patient satisfaction data located on the internet", *Journal of Health Organization and Management* 29(2), pp. 221–233, pMID: 25800334.

Hu, M. & Liu, B. (2004), "Mining and Summarizing Customer Reviews", *in Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM Press, pp. 168–177.

Hu, X., Tang, J., Gao, H. & Liu, H. (2013a), "Unsupervised sentiment analysis with emotional signals", *in 22nd International World Wide Web Conference*, Rio de Janeiro, Brazil: International World Wide Web Conferences Steering Committee / ACM, pp. 607–618.

Hu, X., Tang, L., Tang, J. & Liu, H. (2013b), "Exploiting social relations for sentiment analysis in microblogging", *in Sixth ACM International Conference on Web Search and Data Mining*, Rome, Italy: ACM, pp. 537–546.

Huang, M., Ye, B., Wang, Y., Chen, H., Cheng, J. & Zhu, X. (2014), "New Word Detection for Sentiment Analysis", *in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, MD, USA: ACL, pp. 531–541.

Hunston, S. (2011), *Corpus Approaches to Evaluation: Phraseology and Evaluative Language*, Oxon: Routledge.

Iman, R. L. & Davenport, J. M. (1980), "Approximations of the critical region of the Friedman statistic", *Communications in Statistics* 9(6), pp. 571–595.

Jaggi, M., Uzdilli, F. & Cieliebak, M. (2014), "Swiss-Chocolate: Sentiment Detection using Sparse SVMs and Part-Of-Speech n-Grams", *in Proceedings of the 8th International Workshop on Semantic Evaluation*, Dublin, Ireland: ACL, pp. 601–604.

Jindal, N. & Liu, B. (2006), "Identifying Comparative Sentences in Text Documents", *in Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: ACM Press, pp. 244–251.

Jindal, N. & Liu, B. (2008), "Opinion Spam and Analysis", *in Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, New York, NY, USA: ACM Press, pp. 219–230.

Joachims, T. (1998), "Text Categorization with Suport Vector Machines: Learning with Many Relevant Features", *in Proceedings of the 10th European Conference on Machine Learning*, Chemnitz, Germany: Springer, pp. 137–142.

Joachims, T. (2002), *Learning to classify text using support vector machines: Methods, theory and algorithms*, Kluwer Academic Publishers.

Kamps, J., Marx, M., Mokken, R. J. & De Rijke, M. (2004), "Using WordNet to Measure Semantic Orientations of Adjectives", *in Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisboa, Portugal: ELRA, pp. 1115–1118.

Kanayama, H. & Nasukawa, T. (2006), "Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis", *in Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia: ACL, pp. 355–363.

Karanasou, M., Doulkeridis, C. & Halkidi, M. (2015), "DsUniPi: An SVM-based Approach for Sentiment Analysis of Figurative Language on Twitter", *in Proceedings of the 9th International Workshop on Semantic Evaluation*, Denver, Colorado, USA: ACL, pp. 709–713.

Kessler, B., Numberg, G. & Schütze, H. (1997), "Automatic detection of text genre", *in Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, Spain: ACL, pp. 32–38.

Kilgarriff, A. (1996a), "Comparing word frequencies across corpora: Why chi-square doesnt work, and an improved LOB-Brown comparison", *in Proceedings of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing Conference*, Bergen, Norway.

Kilgarriff, A. (1996b), "Which words are particularly characteristic of a text? A survey of statistical approaches", *in Language Engineering for Document Analysis and Recognition, AISB Workshop Proceedings*, Brighton, England: LEDAR, pp. 33–40.

Kim, S.-M. & Hovy, E. (2004), "Determining the Sentiment of Opinions", *in Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland: COLING, pp. 1367–1373.

Kim, Y. (2014), "Convolutional Neural Networks for Sentence Classification", *in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar: ACL, pp. 1746–1751.

Kinneavy, J. E. (1969), "The Basic Aims of Discourse", *College Composition and Communication* 20(5), pp. 297–304.

Kinneavy, J. L. (1971), *A Theory of Discourse: The Aims of Discourse*, Norton.

Klebanov, B. B., Madnani, N. & Burstein, J. (2013), "Using Pivot-Based Paraphrasing and Sentiment Profiles to Improve a Subjectivity Lexicon for Essay Data", *TACL* 1, pp. 99–110.

Koppel, M. & Schler, J. (2006), "The Importance of Neutral Examples for Learning Sentiment", *Computational Intelligence* 22(2), pp. 100–109.

Landis, J. R. & Koch, G. G. (1977), "The Measurement of Observer Agreement for Categorical Data", *Biometrics* 33(1), pp. 159–174.

LeCun, Y., Bengio, Y. & Hinton, G. (2015), "Deep learning", *Nature* 521(7553), pp. 436–444.

LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998), "Gradient-based learning applied to document recognition", *Proceedings of the IEEE* 86(11), pp. 2278–2324.

Lee, C. S. & Ma, L. (2012), "News sharing in social media: The effect of gratifications and prior experience", *Computers in Human Behavior* 28(2), pp. 331–339.

Leech, G. (1992), "100 million words of English: the British National Corpus (BNC)", *Language Research* 28(1), pp. 1–13.

Lesk, M. (1986), "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone", *in Proceedings of the 5th Annual International Conference on Systems Documentation*, Toronto, Ontario, Canada: ACM, pp. 24–26.

Leskovec, J., Huttenlocher, D. & Kleinberg, J. (2010), "Signed Networks in Social Media", *in Proceedings of the 28th International Conference on Human Factors in Computing Systems*, Atlanta, Georgia, USA: ACM Press, pp. 1361–1370.

Lewis, D. D. (1998), "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval", *in Proceedings of the 10th European Conference on Machine Learning*, Chemnitz, Germany: Springer, pp. 4–15.

Li, S. (2014), *Improving the sentiment classification of stock tweets*, Ph.D. thesis, University of Birmingham.

Li, S., Huang, L., Wang, J. & Zhou, G. (2015), "Semi-Stacking for Semi-supervised Sentiment Classification", *in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, Beijing, China: ACL, pp. 27–31.

Liebrecht, C., Kunneman, F. & van den Bosch, A. (2013), "The perfect solution for detecting sarcasm in tweets #not", *in Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Atlanta, Georgia, USA: ACL, pp. 29–37.

Liu, B. (2010), *Handbook of Natural Language Processing*, Chapman & Hall, chapter Sentiment Analysis and Subjectivity.

Liu, B. (2012), *Sentiment Analysis and Opinion Mining*, Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers.

Liu, B. (2015), *Sentiment Analysis - Mining Opinions, Sentiments, and Emotions*, Cambridge University Press.

Liu, B., Hu, M. & Cheng, J. (2005), "Opinion Observer: Analyzing and Comparing Opinions on the Web", *in Proceedings of the 14th International Conference on World Wide Web*, Chiba, Japan: ACM Press, pp. 342–351.

Liu, Y., Yu, X., Liu, B. & Chen, Z. (2014), "Sentence-Level Sentiment Analysis in the Presence of Modalities", *in Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing*, Kathmandu, Nepal: Springer, pp. 1–16.

Lovins, J. B. (1968), "Development of a stemming algorithm", *Mechanical Translation and Computational Linguistics* 11, pp. 22–31.

Lu, Y., Castellanos, M., Dayal, U. & Zhai, C. (2011), "Automatic construction of a context-aware sentiment lexicon: an optimization approach", *in Proceedings of the 20th International Conference on World Wide Web*, Hyderabad, India: ACM, pp. 347–356.

Luong, T. S., Le, I., Vinyals, Q. & Wojciech, O. Z. (2015), "Addressing the Rare Word Problem in Neural Machine Translation", *in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, China: ACL, pp. 11–19.

Luyckx, K. & Daelemans, W. (2008), "Authorship Attribution and Verification with Many Authors and Limited Data", *in Proceedings of the 22nd International Conference on Computational Linguistics*, Manchester, UK: COLING, pp. 513–520.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y. & Potts, C. (2011), "Learning Word Vectors for Sentiment Analysis", *in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA: ACL, pp. 142–150.

Manning, C. D., Raghavan, P. & Schütze, H. (2008), *Introduction to Information Retrieval*, New York: Cambridge University Press.

Martin, J. R. & White, P. R. (2005), *The Language of Evaluation: Appraisal in English*, London: Palgrave.

Maynard, D. & Greenwood, M. A. (2014), "Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis", *in Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland: ELRA, pp. 4238–4243.

McAuley, J. J. & Leskovec, J. (2013a), "From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews", *in 22nd International World Wide Web Conference*, Rio de Janeiro, Brazil: ACM Press, pp. 897–908.

McAuley, J. J. & Leskovec, J. (2013b), "Hidden factors and hidden topics: understanding rating dimensions with review text", *in Proceedings of the Seventh ACM Conference on Recommender Systems*, Hong Kong, China: ACM, pp. 165–172.

McCallum, A., Nigam, K. et al. (1998), "A Comparison of Event Models for Naive Bayes Text Classification", *in AAAI-98 Workshop on Learning for Text Categorization*, AAAI, volume 752, pp. 41–48.

McCallum, A. K. (1996), *Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering*, available from `http://www.cs.cmu.edu/~mccallum/bow`.

McEnery, T. & Hardie, A. (2011), *Corpus linguistics: Method, theory and practice*, Cambridge University Press.

McNemar, Q. (1947), "Note on the sampling error of the difference between correlated proportions or percentages", *Psychometrika* 12(2), pp. 153–157.

Mejova, Y. & Srinivasan, P. (2012), "Crossing Media Streams with Sentiment: Domain Adaptation in Blogs, Reviews and Twitter", *in Proceedings of the Sixth International Conference on Weblogs and Social Media*, Dublin, Ireland: The AAAI Press, pp. 234–241.

Melzi, S., Abdaoui, A., Azé, J., Bringay, S., Poncelet, P. & Galtier, F. (2014), "Patient's rationale: Patient Knowledge retrieval from health forums", *in eTELEMED'2014: 6th International Conference on eHealth, Telemedicine, and Social Medicine*, Barcelona, Spain, pp. 140–145.

Mihalcea, R., Corley, C. & Strapparava, C. (2006), "Corpus-based and Knowledge-based Measures of Text Semantic Similarity", *in Proceedings of the Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference*, Boston, Massachusetts, USA: AAAI Press, pp. 775–780.

Mihalcea, R. & Moldovan, D. I. (1999), "A Method for Word Sense Disambiguation of Unrestricted Text", *in Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Maryland, USA: ACL, pp. 152–158.

Miller, G. A. (1995), "WordNet: A Lexical Database for English", *Communications of the ACM* 38(11), pp. 39–41.

Miller, G. R. (2002), *The Persuasion Handbook: Developments in Theory and Practice*, Sage.

Miller, M., Sathi, C., Wiesenthal, D., Leskovec, J. & Potts, C. (2011), "Sentiment Flow Through Hyperlink Networks", *in Proceedings of the Fifth International Conference on Weblogs and Social Media*, Barcelona, Spain: AAAI Press, pp. 550–553.

Mohammad, Saifand Turney, P. (2010), "Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon", *in Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles, CA, USA: ACL, pp. 26–34.

Mohammad, S., Kiritchenko, S. & Zhu, X. (2013), "NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets", *in Proceedings of the 7th International Workshop on Semantic Evaluation*, Atlanta, Georgia, USA: ACL, pp. 321–327.

Mohammad, S. M., Salameh, M. & Kiritchenko, S. (2015), "How translation alters sentiment", *Journal of Artificial Intelligence Research* 1, pp. 1–20.

Montoyo, A., Suárez, A., Rigau, G. & Palomar, M. (2005), "Combining Knowledge- and Corpus-based Word-Sense-Disambiguation Methods", *Journal of Artificial Intelligence Research* 23, pp. 299–330.

Morency, L., Mihalcea, R. & Doshi, P. (2011), "Towards multimodal sentiment analysis: harvesting opinions from the web", *in Proceedings of the 13th International Conference on Multimodal Interfaces*, Alicante, Spain: ACM, pp. 169–176.

Mukherjee, A. (2014), *Probabilistic Models for Fine-Grained Opinion Mining: Algorithms and Applications*, Ph.D. thesis, University of Illinois at Chicago.

Mukherjee, A. & Liu, B. (2012), "Modeling Review Comments", *in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju Island, Korea: ACL, pp. 320–329.

Mukherjee, A. & Liu, B. (2013), "Discovering User Interactions in Ideological Discussions", *in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria: ACL, pp. 671–681.

Mukras, R. (2009), *Representation and Learning Schemes for Sentiment Analysis*, Ph.D. thesis, Robert Gordon University.

Mullen, T. & Collier, N. (2004), "Sentiment Analysis using Support Vector Machines with Diverse Information Sources", *in Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain: ACL, pp. 412–418.

Murakami, A. & Raymond, R. (2010), "Support or Oppose?: Classifying Positions in Online Debates from Reply Activities and Opinion Expressions", *in Proceedings of the 23rd International Conference on Computational Linguistics: Posters Volume*, Beijing, China: Chinese Information Processing Society of China, pp. 869–875.

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A. & Wilson, T. (2013), "SemEval-2013 Task 2: Sentiment Analysis in Twitter", *in Proceedings of the 7th International Workshop on Semantic Evaluation*, Atlanta, Georgia, USA: ACL, pp. 312–320.

Nalisnick, E. & Baird, H. (2013), "Character-to-Character Sentiment Analysis in Shakespeares Plays", *in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Sofia, Bulgaria: ACL, pp. 479–483.

Nemenyi, P. B. (1963), *Distribution-free Multiple Comparisons*, Ph.D. thesis, Princeton University.

Ng, A. Y. & Jordan, M. I. (2001), "On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes", *in Advances in Neural Information Processing Systems*, Vancouver, British Columbia, Canada: MIT Press, pp. 841–848.

Nguyen, Q. D., Nguyen, Q. D. & Pham, B. S. (2013), "A Two-Stage Classifier for Sentiment Analysis", *in Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Nagoya, Japan: Asian Federation of Natural Language Processing / ACL, pp. 897–901.

Nguyen, T. H. & Shirai, K. (2015), "Topic Modeling based Sentiment Analysis on Social Media for Stock Market Prediction", *in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, Beijing, China: ACL, pp. 1354–1364.

NHS England (2014), "NHS England Review of the Friends and Family Test", `https://www.england.nhs.uk/wp-content/uploads/2014/07/fft-rev.pdf`, last accessed: 24-11-2015.

NHS England (2015), "5 millionth feedback landmark for Friends and Family Test", `https://www.england.nhs.uk/2015/02/09/five-million/`, last accessed: 24-11-2015.

Niu, Y., Zhu, X., Li, J. & Hirst, G. (2005), "Analysis of Polarity Information in Medical Text", *in Proceedings of the American Medical Informatics Association 2005 Annual Symposium*, Washington D.C, USA: AMIA, pp. 570–574.

O'Connor, B., Balasubramanyan, R., Routledge, B. R. & Smith, N. A. (2010), "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series", *in Proceedings of the Fourth International Conference on Weblogs and Social Media*, Washington, DC, USA: The AAAI Press.

Ofek, N., Caragea, C., Rokach, L., Biyani, P., Mitra, P., Yen, J., Portier, K. & Greer, G. (2013), "Improving Sentiment Analysis in an Online Cancer Survivor Community Using Dynamic Sentiment Lexicon", *in International Conference on Social Intelligence and Technology*, pp. 109–113.

Ortony, A., Clore, G. L. & Collins, A. (1988), *The Cognitive Structure of Emotions*, Cambridge: Cambridge University Press.

Palanisamy, P., Yadav, V. & Elchuri, H. (2013), "Serendio: Simple and Practical lexicon based approach to Sentiment Analysis", *in proceedings of Second Joint Conference on Lexical and Computational Semantics*, pp. 543–548.

Paltoglou, G. & Thelwall, M. (2010), "A Study of Information Retrieval Weighting Schemes for Sentiment Analysis", *in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden: ACL, pp. 1386–1395.

Pan, S. J. & Yang, Q. (2010), "A Survey on Transfer Learning", *IEEE Transactions on Knowledge and Data Engineering* 22(10), pp. 1345–1359.

Pang, B. & Lee, L. (2004), "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts", *in Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain: ACL, pp. 271–278.

Pang, B. & Lee, L. (2005), "Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales", *in Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, University of Michigan, USA: ACL, pp. 115–124.

Pang, B. & Lee, L. (2008), "Opinion mining and sentiment analysis", *Foundations and trends in information retrieval* 2(1-2), pp. 1–135.

Pang, B., Lee, L. & Vaithyanathan, S. (2002), "Thumbs up?: Sentiment Classification Using Machine Learning Techniques", *in Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA, USA: ACL, pp. 79–86.

Papakonstantinou, P. A., Xu, J. & Cao, Z. (2014), "Bagging by Design (on the Suboptimality of Bagging)", *in Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, Québec City, Québec, Canada: AAAI Press, pp. 2041–2047.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A. & Booth, R. J. (2007), *The Development and Psychometric Properties of LIWC2007*, Austin, Texas, USA, available from `http://www.liwc.net`.

Pestian, J. P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., Cohen, K. B., Hurdle, J. & Brew, C. (2012), "Sentiment Analysis of Suicide Notes: A Shared Task", *Biomedical Informatics Insights* 5(1), pp. 3–16.

Picker Institute (2015), "Using Patient Feedback", `http://www.nhssurveys.org/Filestore/documents/QIFull.pdf`, last accessed: 24-11-2015.

Pilehvar, M. T., Jurgens, D. & Navigli, R. (2013), "Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity", *in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria: ACL, pp. 1341–1351.

Platt, J. (1999), "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods", *in Advances in Large Margin Classifiers*, volume 10, pp. 61–74.

Polanyi, L. & Zaenen, A. (2006), "Contextual valence shifters", *in Computing Attitude and Affect in Text: Theory and Applications*, Dordrecht: Springer, pp. 1–10.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S. & Androutsopoulos, I. (2015), "SemEval-2015 Task 12: Aspect Based Sentiment Analysis", *in Proceedings of the 9th International Workshop on Semantic Evaluation*, Denver, Colorado: ACL, pp. 486–495.

Poria, S., Cambria, E. & Gelbukh, A. F. (2015), "Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-level Multimodal Sentiment Analysis", *in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisboa, Portugal: ACL, pp. 2539–2544.

Portier, K., Greer, G. E., Rokach, L., Ofek, N., Wang, Y., Biyani, P., Yu, M., Banerjee, S., Zhao, K., Mitra, P. et al. (2013), "Understanding topics and sentiment in an online cancer survivor community", *Journal of the National Cancer Institute* 47, pp. 195–198.

Prechelt, L., Malpohl, G. & Philippsen, M. (2000), "JPlag: Finding plagiarisms among a set of programs", Technical Report 2000-1, Karlsruhe Institute of Technology.

Proserpio, D. & Zervas, G. (2014), "Online Reputation Management: Estimating the Impact of Management Responses on Consumer Reviews", Technical Report Research Paper No. 2521190, Boston U. School of Management.

Pustejovsky, J. & Stubbs, A. (2012), *Natural Language Annotation for Machine Learning*, O'Reilly Publishers.

Qiu, B., Zhao, K., Mitra, P., Wu, D., Caragea, C., Yen, J., Greer, G. E. & Portier, K. (2011), "Get Online Support, Feel Better - Sentiment Analysis and Dynamics in an Online Cancer Survivor Community", *in Proceedings of the Third International Conference on Social Computing*, Boston, MA, USA: IEEE, pp. 274–281.

Rayson, P. & Garside, R. (2000), "Comparing Corpora using Frequency Profiling", *in Proceedings of the Workshop on Comparing Corpora*, Hong Kong, China: ACL, pp. 1–6.

Read, J. (2005), "Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification", *in Proceedings of the ACL Student Research Workshop*, ACLstudent '05, Stroudsburg, PA, USA: ACL, pp. 43–48.

Rifkin, R. & Klautau, A. (2004), "In Defense of One-Vs-All Classification", *Journal of Machine Learning Research* 5, pp. 101–141.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N. & Huang, R. (2013), "Sarcasm as Contrast between a Positive Sentiment and Negative Situation", *in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA: ACL, pp. 704–714.

Riloff, E. & Wiebe, J. (2003), "Learning Extraction Patterns for Subjective Expressions", *in Proceedings of the 2003 conference on Empirical Methods in Natural Language Processing*, Sapporo, Japan: ACL, pp. 105–112.

Rokach, L. (2010), "Ensemble-based classifiers", *Artificial Intelligence Review* 33(1-2), pp. 1–39.

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A. & Stoyanov, V. (2015), "SemEval-2015 Task 10: Sentiment Analysis in Twitter", *in Proceedings of the 9th International Workshop on Semantic Evaluation*, Denver, Colorado, USA: ACL, pp. 451–463.

Sadamitsu, K. & Yamamoto, M. (2008), "Sentiment Analysis Based on Probabilistic Models Using Inter-Sentence Information", *in Proceedings of the Sixth International Language Resources and Evaluation*, Marrakech, Morocco: ELRA, pp. 2892–2896.

Salah, Z. I. S. (2014), *Machine Learning and Sentiment Analysis Approaches for the Analysis of Parliamentary Debates*, Ph.D. thesis, University of Liverpool, UK.

Sarker, A., Molla, D. & Paris, C. (2011), "Outcome Polarity Identification of Medical Papers", *in Proceedings of the Australasian Language Technology Association Workshop 2011*, Canberra, Australia: ACL, pp. 105–114.

Scheible, C. & Schütze, H. (2013), "Sentiment Relevance", *in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria: ACL, pp. 954–963.

Scherer, K. R. (1984), "Emotion as a multicomponent process: A model and some cross-cultural data", *in* Shaver, P. (Ed.), *Review of Personality and Social Psychology: Emotions, Relationships and Health*, Beverley Hills, CA, USA: Sage Publications, Inc., pp. 37–63.

Schmid, H. (1994), "Probabilistic Part-of-Speech Tagging Using Decision Trees", *in Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Schneider, A. & Dragut, E. C. (2015), "Towards Debugging Sentiment Lexicons", *in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, Beijing, China: ACL, pp. 1024–1034.

Schneider, K. (2004), "On Word Frequency Information and Negative Evidence in Naive Bayes Text Classification", *in Proceedings of the 4th International Conference Advances in Natural Language Processing*, Alicante, Spain: Springer, pp. 474–486.

Schulder, M. & Hovy, E. (2014), "Metaphor Detection through Term Relevance", *in Proceedings of the Second Workshop on Metaphor in NLP*, Baltimore, MD, USA: ACL, pp. 18–26.

Searle, J. R. (1976), "A classification of illocutionary acts", *Language in Society* 5(01), pp. 1–23.

Sebastiani, F. (2002), "Machine Learning in Automated Text Categorization", *ACM Computing Surveys* 34(1), pp. 1 – 47.

Shannon, C. E. (1948), "A Mathematical Theory of Communication", *Bell Systems Technical Journal* 27(3), pp. 379–423.

Sharif, H., Zaffar, F., Abbasi, A. & Zimbra, D. (2014), "Detecting adverse drug reactions using a sentiment classification framework", *in Proceedings of the Sixth International Conference on Social Computing*, Stanford University: ASE, pp. 1–10.

Sharma, R., Gupta, M., Agarwal, A. & Bhattacharyya, P. (2015), "Adjective Intensity and Sentiment Analysis", *in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal: ACL, pp. 2520–2526.

Simančík, F. & Lee, M. (2009), "A CCG-based system for valence shifting for sentiment analysis", *in Advances in Computational Linguistics: Proceedings of 10th International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico, volume 41, pp. 93–102.

Sinclair, J. (1999), "A way with common words", *Language and Computers* 26, pp. 157–180.

Smailović, J., Grčar, M., Lavrač, N. & Žnidaršič, M. (2014), "Stream-based active learning for sentiment analysis in the financial domain", *Information Sciences* 285, pp. 181–203.

Smith, P. & Lee, M. (2012), "A CCG-Based Approach to Fine-Grained Sentiment Analysis", *in Proceedings of the 2nd Workshop on Sentiment Analysis where AI meets Psychology*, Mumbai, India: The COLING 2012 Organising Committee, pp. 3–16.

Snyder, B. & Barzilay, R. (2007), "Multiple Aspect Ranking Using the Good Grief Algorithm", *in Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, Rochester, New York, USA: ACL, pp. 300–307.

Socher, R., Huval, B., Manning, C. D. & Ng, A. Y. (2012), "Semantic Compositionality through Recursive Matrix-Vector Spaces", *in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea: ACL, pp. 1201–1211.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, D. C., Ng, A. & Potts, C. (2013), "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank", *in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA: ACL, pp. 1631–1642.

Sokolova, M. & Bobicev, V. (2013), "What Sentiments Can Be Found in Medical Forums?", *in Proceedings of 9th International Conference on the Recent Advances in Natural Language Processing*, Hissar, Bulgaria: ACL, pp. 633–639.

Somasundaran, S. (2010), *Discourse-level Relations for Opinion Analysis*, Ph.D. thesis, University of Pittsburgh.

Somasundaran, S., Ruppenhofer, J. & Wiebe, J. (2007), "Detecting Arguing and Sentiment in Meetings", *in Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium: ACL, pp. 26–34.

Stone, P. J. (1966), *The General Inquirer: A Computer Approach to Content Analysis*, The MIT Press.

Strapparava, C. & Mihalcea, R. (2007), "SemEval-2007 Task 14: Affective Text", *in Proceedings of the 4th International Workshop on Semantic Evaluations*, Prague, Czech Republic: ACL, pp. 70–74.

Strzalkowski, T. (1995), "Natural Language Information Retrieval", *Information Processing & Management* 31(3), pp. 397–417.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & Stede, M. (2011), "Lexicon-Based Methods for Sentiment Analysis", *Computational Linguistics* 37(2), pp. 267–307.

Talbot, R., Acheampong, C. & Wicentowski, R. (2015), "SWASH: A Naive Bayes Classifier for Tweet Sentiment Identification", *in Proceedings of the 9th International Workshop on Semantic Evaluation*, Denver, Colorado, USA: ACL, pp. 626–630.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. & Kappas, A. (2010), "Sentiment Strength Detection in Short Informal Text", *Journal of the American Society for Information Science and Technology* 61(12), pp. 2544–2558.

Thomas, M., Pang, B. & Lee, L. (2006), "Get out the Vote: Determining Support or Opposition from Congressional Floor-debate Transcripts", *in Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia: ACL, pp. 327–335.

Trigg, L. (2011), "Patients' opinions of health care providers for supporting choice and quality improvement", *Journal of Health Services Research & Policy* 16(2), pp. 102–107.

TripAdvisor (2014), "How to add Management Responses to TripAdvisor Traveller Reviews", `https://www.tripadvisor.co.uk/TripAdvisorInsights/n2428/how-add-management-responses-tripadvisor-traveller-reviews`, last accessed: 24-11-2015.

Turney, P. D. (2002), "Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classificationl of Reviews", *in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA: ACL, pp. 417–424.

Uzdilli, F., Jaggi, M., Egger, D., Julmy, P., Derczynski, L. & Cieliebak, M. (2015), "Swiss-Chocolate: Combining Flipout Regularization and Random Forests with Artificially Built Subsystems to Boost Text-Classification for Sentiment", *in Proceedings of the 9th International Workshop on Semantic Evaluation*, Denver, Colorado, USA,: ACL, pp. 608–612.

Vapnik, V. (1995), *The Nature of Statistical Learning Theory*, New York: Springer-Verlag.

Velikovich, L., Blair-Goldensohn, S., Hannan, K. & McDonald, R. T. (2010), "The viability of web-derived polarity lexicons", *in Proceedings of the Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, Los Angeles, California, USA: ACL, pp. 777–785.

Volkova, S., Wilson, T. & Yarowsky, D. (2013), "Exploring Sentiment in Social Media: Bootstrapping Subjectivity Clues from Multilingual Twitter Streams", *in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria: ACL, pp. 505–510.

Wan, X. (2009), "Co-training for Cross-lingual Sentiment Classification", *in Proceedings of the 47th Annual Meeting of the Associatio for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Singapore: ACL, pp. 235–243.

Wang, S. & Manning, C. D. (2012), "Baselines and bigrams: Simple, good sentiment and topic classification", *in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju Island, Korea: ACL, pp. 90–94.

West, R., Paskov, S. H., Leskovec, J. & Potts, C. (2014), "Exploiting Social Network Structure for Person-to-Person Sentiment Analysis", *Transactions of the Association for Computational Linguistics* 2(1), pp. 297–310.

Whitelaw, C., Garg, N. & Argamon, S. (2005), "Using Appraisal Groups for Sentiment Analysis", *in Proceedings of the 2005 International Conference on Informationand Knowledge Management*, Bremen, Germany: ACM Press, pp. 625–631.

Wicentowski, R. (2015), "SWATCS65: Sentiment Classification Using an Ensemble of Class Projects", *in Proceedings of the 9th International Workshop on Semantic Evaluation*, Denver, Colorado, USA: ACL, pp. 631–635.

Wiebe, J., Wilson, T., Bruce, R., Bell, M. & Martin, M. (2004), "Learning Subjective Language", *Computational Linguistics* 30(3), pp. 277–308.

Wiebe, J. M., Bruce, R. F. & O'Hara, T. P. (1999), "Development and Use of a Gold-standard Data Set for Subjectivity Classifications", *in Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Maryland, USA: ACL, pp. 246–253.

Wilks, Y. (1990), "Providing machine tractable dictionary tools", *Machine Translation* 2, pp. 341–401.

Wilson, T. (2008a), "Annotating Subjective Content in Meetings.", *in Proceedings of the Sixth International Language Resources and Evaluation (LREC08)*, Marrakech, Morocco.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E. & Patwardhan, S. (2005a), "OpinionFinder: A System for Subjectivity Analysis", *in Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada: ACL, pp. 34–35.

Wilson, T. & Wiebe, J. (2005), "Annotating Attributions and Private States", *in Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, Ann Arbor, Michigan: ACL, pp. 53–60.

Wilson, T., Wiebe, J. & Hoffmann, P. (2005b), "Recognizing Contextual Polarity in Phrase-level Sentiment Analysis", *in Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada: ACL, pp. 347–354.

Wilson, T., Wiebe, J. & Hwa, R. (2004), "Just How Mad Are You? Finding Strong and Weak Opinion Clauses", *in Proceedings of the 19th National Conference on Artifical Intelligence*, San Jose, California, USA: AAAI Press / The MIT Press, pp. 761–767.

Wilson, T. A. (2008b), *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intenity, Polarity, and Attitudes of Private States*, University of Pittsburgh.

Wise, M. J. (1993), "String similarity via greedy string tiling and running Karp-Rabin matching", *Online Preprint, Dec* 119.

Wittgenstein, L. (1953), *Philosophical Investigations. (Translated by Anscombe, G.E.M.)*, Oxford: Basil Blackwell Ltd.

Wolpert, D. H. (1992), "Stacked generalization", *Neural Networks* 5(2), pp. 241–259.

Xia, L., Gentile, A. L., Munro, J. & Iria, J. (2009), "Improving Patient Opinion Mining Through Multi-step Classification", *in Text, Speech and Dialogue, 12th International Conference*, Pilsen, Czech Republic: Springer, pp. 70–76.

Xia, R., Zong, C. & Li, S. (2011), "Ensemble of feature sets and classification algorithms for sentiment classification", *Information Sciences* 181(6), pp. 1138–1152.

Yang, H., Willis, A., De Roeck, A. & Nuseibeh, B. (2012), "A hybrid model for automatic emotion recognition in suicide notes", *Biomedical informatics insights* 5(Suppl 1), p. 17.

Yih, W., He, X. & Meek, C. (2014), "Semantic Parsing for Single-Relation Question Answering", *in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, MD, USA: ACL, pp. 643–648.

Zadrozny, B. & Elkan, C. (2001), "Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers", *in Proceedings of the Eighteenth International Conference on Machine Learning*, Williamstown, MA, USA: Morgan Kaufmann, pp. 609–616.

Zhang, K., Cheng, Y., Xie, Y., Honbo, D., Agrawal, A., Palsetia, D., Lee, K., Liao, W.-k. & Choudhary, A. (2011), "SES: Sentiment elicitation system for social media data", *in Proceedings of the 11th International Conference on Data Mining Workshops*, Vancouver, BC, Canada: IEEE, pp. 129–136.

Zhang, L. & Liu, B. (2011), "Identifying Noun Product Features that Imply Opinions", *in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA: ACL, pp. 575–580.

Zhang, Y., Zhang, H., Zhang, M., Liu, Y. & Ma, S. (2014), "Do users rate or review?: boost phrase-level sentiment labeling with review-level sentiment classification", *in Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Gold Coast , QLD, Australia: ACM, pp. 1027–1030.

Zhao, K., Yen, J., Greer, G., Qiu, B., Mitra, P. & Portier, K. (2014), "Finding influential users of online health communities: a new metric based on sentiment influence", *Journal of the American Medical Informatics Association* 21(1), pp. 212–218.

Zhou, S., Chen, Q., Wang, X. & Li, X. (2014), "Hybrid Deep Belief Networks for Semi-supervised Sentiment Classification", *in Proceedings of the 25th International Conference on Computational Linguistics*, Dublin, Ireland: ACL, pp. 1341–1349.